Human Pose Estimation and Activity Recognition Using Deep Learning for Smart Cities

Adnan Raghib, Shahbaz Ahmad, Md Imran Hussain

B.Tech Computer Science And
Engineering (2022 Batch)Lovely
Professional University, Jalandhar

Abstract

In machine learning, a convolutional neural network (CNN, or ConvNet) is a class of deep, feed-forward artificial neural networks, most commonly applied to analyzing visual imagery. CNNs use a variation of multilayer perceptrons designed to require minimal preprocessing. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics. Convolutional networks were inspired by biological processes in that the connectivity pattern between neurons resembles the organization of the animal visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field. CNNs use relatively little pre-processing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms were hand-engineered. This independence from prior knowledge and human effort in feature design is a major advantage. They have applications in image and video recognition, recommender systems and natural language processing**.** When you combine them you get great results for complex problems like computer vision video classification to implement human action recognition on videos. For pose estimationwe use Mediapipe. MediaPipe Pose is an Machine Learning solution that tracks body pose with precision using 33 landmarks. It utilizes BlazePose research by removing background from a complete RGB frame. Pose- generated pipelines are lightweight opposing conventional ML solutions.

1 Introduction

To perform human activity recognition we have to dealwith videos which is a bit tricky than our usual image recognition. Videos are nothing but sequences and framesof different images. But if we think of activityrecognition(video) same as image recognition then we willget different results for the different frames of the images.In this paper we will be using different architectures andapproaches to train our model so that it can predict bestresults on our unseen video data. At last we will pick thebest performing model and deploy it in our web app. Poseestimation is also a major part of research today. Its application varies from various areas of today's modernworld. Some examples are sports, healthcare, virtualenvironment reality for animation and entertainment. [1, 2,3]. Apart from these the posture detection can be used indifferent spectrum of human lives and healthcare.Increasing number of elderly population is always awithout sufficient healthcare resource is always a challengefor any country. So it is really important to have atechnology which can support and monitor their activitiesremotely which will give them a sense of less vulnerabilityand they can live more independently. [11, 12, 13]. Weshould always maintain a healthy posture to live a healthylife. Posture is all about the way people hold their limbsand position their body. Since the boom in technologyhuman have developed a lazy lifestyle which leads to lessphysical activity and movement. [14, 15, 16, 17, 18, 19].Sitting continuously at a single place for a very long timeeither for studying or work leads to decrease in musclestrength. This unhealthy life style which we have adoptedand not paying much attention about the body posturealways leads to pain in neck back and shoulder. Therefore it is really important to maintain our posture during work and study.

Considering the need, the paper reports three major contributions that are outlined below:

1. In this paper, we tried to implemented an architecture which can automaticaly identify the human activity that is being performed in the video with a high range of accuracy. For good user experience we have also created a web user interface to and deployed our model in backend to check accuracy of any given inputvideo.
2. We have also used mediapipe and OpenCV toimplement a pose estimation model which canmultiple use cases in health and sports.
3. We have analysed results of our different approachesdeeply and plotted the accuracy table to evaluate thebest performing model for our purpose.

The paper mainly discuss deep learning approach to experiment with our dataset which includes different videos of different classes to recognise the human activity in the video. We tried to implement CNN first and found good results but when we merge it with LSTM the result were much more accurate.

The paper id divided into 5 sections. In the coming sectionwe have discussed the work within the same area of study and research. In section 3 we discussed about how we implemented the different approaches for our solution. In section 4 we have the results of our different methods which we have analysed and discussed thoroughly. Section5 provides the conclusion of the paper.

2. Related Work

In the research field a very vast work has been done about human activity recognition and pose estimation because these has a very large use cases. In this section we will some of these works.

In supervision of human health technology and portable system is must. Thus in this paper posture recognition is

determined using Long Range Technology (LoRa). As the name suggests it has the advantage of very long transmission at an affordable and low cost. Wearable cloths are designed using these two technology musisensory and LoRa which provides comfort in any posture. Multiprocessing is preferred in this paper because it has high transmitting frequency and data size is large. Sliding window is used to do multiprocessing and for data preprocessing and feature extraction Random Forest is used [20]. Human posture is also a way of communication but it is non-verbal. In this paper by Iulian Radu, Ethan Tu, and Bertrand Schneider [21], they have used augmented reality to detect the human posture with really advanced body and muscle tracking technology. This method is also very cost friendly. Some machine learning algorithms on moving body posture is also used to find group coperation [21]. Posture play a very important part to perform physical exercise such as yoga. So it is a very important use case but the problem is that we have a very less data set as well as it involves real time bases which is really challenging. So to overcome this challenge in this paper more than 5500 images of variety of yoga poses is captured [22, 23]. Tf pose estimation algorithm is very popular for detection of pose on real time human body. Different angles of body is detected by tf-pose skeleton which are captured and can be used by machine learning algorithms. We can use different available algorithms to determine it but the best performing model mong all is Random forest. It gives the best accuracy. [22, 23, 24].

The modern work and daily life style of human has forced us to spend much of our day sitting. This is yet another problem for developing poor posture of our body which effects both physical and mental health. In this paper data set is designed for effective sitting and stretch position. After that smart cushion and other Artificial Intelligence technology plus pressure sensing is used to determine the pose recognition. Different machine learning models are trained for more than 13 different pose which gives amazing performances [25]. In another paper, a unique type of chair is made with sensors which tries to avoid wrong sitting position which may cause harm. Random forest and decision tree are used to compare the analysis. Random forest gives the better result and thus was preferred [26]. Sitting Posture Monitoring Systems are used to improve the sitting posture (SPMSs). It has fixed sensors on different parts of the chair. Six different pose are considered for this experiment. Different machine learning and deep learning models are applied on the fetched data with body weight ratio measured by SPMS. Support Vector Machine (SVM) gives better result as compared to others [27]. So far we haven't considered the case for specially abled people. In the next paper these is an algorithm

designed to detect posture of person sitting on wheel chair. CNN type neighbourhood rule is used to select data from network of sensors, then Kennard-stone algorithm is used for data balancing and dimensionality reduction by principal component analysis also called PCA. At last KNN (k-nearest-neighbour) algorithm is applied to our reduced and balanced data and the results were remarkable [28].

A habit once formed can't be changed easily. Similar is the case with Postural Habit, so it is advicable by experts and doctors to develop and form good and healthy posture habit since childhood. Therefore in this paper different machine learning algorithms are applied for posture detection and correction in children. A cushion with sensors is developed and put inside children seat to collect their data. More than 10 children participated for different postures. The highest accuracy found was of Conv Neural Network (CNN) compared to other algorithms[29]. Dance is an art of posing your body which is also a use case in pose recognition. It is but a challenging task because it has different forms and duration time or space is not predefined. Thus for its analysis few things must be known prior like the form of dance. This paper mainly focusses on Classical Indian Dance also called

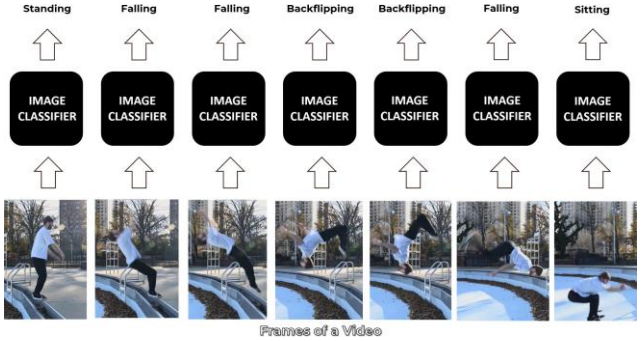
Bharatnatyam. The sequence of music and steps are collected as dataset. After that recognition is performed using different algorithms and models. HMM is applied for sequence recognition. The best performing classifier in this case was also CNN [30].

3. Methodology

This section deals with our method and approach towards developing high accuracy activity recognition model. We have used different approaches and have tested their accuracy with different metrics. We have talked about all these in following subsection.

SINGLE FRAME CLASSIFICATION

The most basic way of classifying any video would be to classify each independent frames individually and store the output, then choose the majority of these output and say it is the final result. We have already seen so many classification techniques for images which performs really well. Most of them trained on deep learning models and CNNs. But for videos we have to take account for different frames. Suppose implementing this approach for a video in



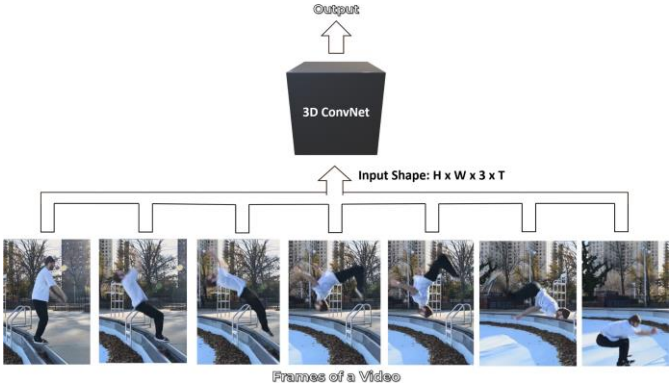
which a boy is doing backflipping so we might get result something like this.

The model predicts falling and other classes in some frames because this approached doesn't care about the sequence of data. Not just the model if some person also looked at these frames differently he might think he is falling. The most simple approach to get the final result is to choose the majority among the predicted class, which will be right for simple cases but in our case falling in not right. Another way might be to get an average of probabilities of predictions and get a more accurate result.

But this method is not accurate because it does not takes account of sequence of data.

USING 3D CNN'S (AKA SLOWFUSION)

A Other option is to use 3D CNN or Slow Fusion where temporary and continuous information are merged together throughout the network which is also why it is called slow fusion. But it is really expensive computational wise and thus not a effective approach for our solution



A Convolutional Neural Network (CNN) is deep learning network specially made to work with images. It considers the pixels of image as a matrix of numbers and max/min pooling on it in a network. It works

exceptionally well with image data and is good for any type of classification involving images. It works with filters which go over image and determine whether a particular feature is present on this part of image or not. The features found are stored in a feature map and as we go deeper in network the number of these map increase and size decrease because of pooling.

As J.K. Aggarwal, Lu Xia mentioned in their paper [10], there are so many ways to generate the motion data using different types of markers at different places in the body like hands bone joints boney regions etc. We can use multiple cameras to estimate the position of all these markers on the body. So all the motion that happens by these body part over the day is recorded thus it can be called as motion capture data.

USING LSTM

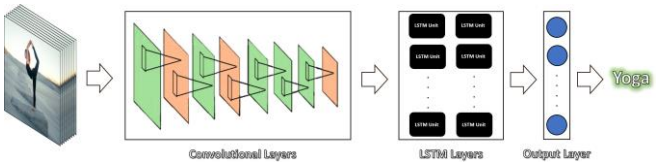
We can use landmark coordinates of the human present in the video to detect a pose detection network for each frame in the video. We can then provide these landmarks as an input to the LSTM network and then it can predict the human activity. At present there are already so much accurate and efficient pose detectors available which can be used for this purpose. But the negative point with this approach is that we have to discard all the other information available in the video like background environment and all. These other informations sometimes can be very useful like in a video if someone is playing football then the the presence of ground and ball and the uniform can play a crucial role in the determining the activity. But with this approach all those other things are ignored.

LSTM is designed for the continuous data and to take previous data into consideration before performing any output. It takes care of the sequence of data. It is actually a type or a part of neural network (NN) called recurrent neural network (RNN), but recurrent neural network are not known for these type of solutions. They are not effective in dealing with long input sequence or long term

dependency because of a problem called vanishing gradient problem in them. LSTMs are make to avoid this problem of vanishing gradient so that it can remember long sequence of inputs. This makes it super powerful while dealing with sequential data which might involve predictions dependent on time like speech recognition, language translation, or music composition. In this paper we will only use LSTM in developing better model for recognition of human activity in the videos.

COMBINING LSTM With CNN

We In this method we will implement LRCN along with something else called LSTM layers in a single model. Other similar approaches might be using a CNN model and LSTM model trained separately. By using both we can take advantage of both of these methods like CNN can be used to extract features from different frames of the video while LSTM can then use these features extracted by CNN to predict actions performed in the video by taking account of the previous inputs as well. But in this approach we will implement something different known as the Long-term Recurrent Convolutional Network (LRCN), which combines CNN and LSTM layers in a single model. To extract spatial features from the frames convolutional layers are used, LSTM is then fed with these extracted spatial features at each time steps for temporal sequence modeling. This results in a robust model which is learned by spatiotemporal features directly in an end to end training.

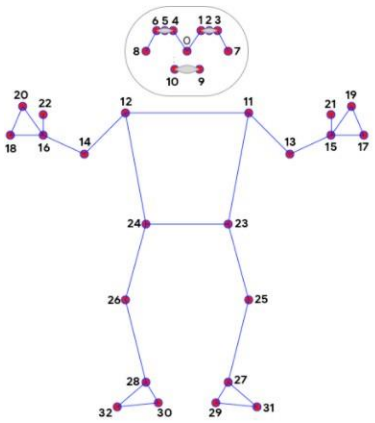


In this paper Jeff Donahue (CVPR 2015) called Long-term Recurrent Convolutional Networks for Visual Recognition and Description, he used this technique [11].

For every frame of this video we will use time distributed wrapper layer which allows applying same layer to every frame independently. This is done so that it makes a layer around which it is wrapped. This makes it capable of taking inputs of shape like number_of_frames, width, height, number_of_channels. This is very necessary because it allows to input the whole video into the model in a single shot, otherwise it would have been very difficult for taking the input to model in a single shot.

Pose Estimation Using OpenCV and MediaPipe

Pose takes in two arguments: detection confidence and tracking confidence. These are simply the strictness of preciseness we want in our resulting pipeline. The pipeline processes the RGB frame and produces an array of landmark coordinates. Each frame updates the landmark coordinates. OpenCV library provides a built-in solution to engage a streaming device, capture a video stream, and provide video frames. We can use this by calling the OpenCV VideoCapture library. This library can read frames of video and display them in a window. The frames extracted from OpenCV are BGR format. So, we first convert it to RGB format. Once we have our video frames in RGB, we can apply MediaPipe's Pose on video frames to track body posture.



0. nose
1. left_eye_inner
2. left_eye
3. left_eye_outer
4. right_eye_inner
5. right_eye
6. right_eye_outer
7. left_ear
8. right_ear
9. mouth_left
10. mouth_right
11. left_shoulder
12. right_shoulder
13. left_elbow
14. right_elbow
15. left_wrist
16. right_wrist
17. left_pinky
18. right_pinky
19. left_index
20. right_index
21. left_thumb
22. right_thumb
23. left_hip
24. right_hip
25. left_knee
26. right_knee
27. left_ankle
28. right_ankle
29. left_heel
30. right_heel
31. left_foot_index
32. right_foot_index

Humans pose extracted from videos can play a very important role in different aspects of modern lifestyle. The various applications like sign language recognition full body posture control physical exercises, pose while doing sports activities, these all require some kind of monitoring. It can form virtual yoga dance fitness classes online by recognising your pose through a camera. In augmentedreality this information on top of physical world can enableoverlay of digital content. MediaPipe is an excellent pertained model to determine

the pose of person either by live actions using camera or by recorded videos. It gives a very accurate and high performing results.

4. EXPERIMENTAL RESULTS

Different classifiers like 2D-CNN and LSTMs are trained to classify different activities as we have seen in theabove section. We extracted different features from ourdata set which was UCF50 - Activity Recognition Dataset, which consists of different videos for different actions seperately stored in different folders. These videos are downloaded from YouTube and are of average size. (doesn't contain sounds to decrease video size) The Datasetcontains:

i. **50** categories of action containing human
ii. **25** video groups are present for each of the actioncategory
iii. **133** on average video persent for each category
iv. **199** no of frames in each video in dataset
v. **320** width of frame average per video
vi. **240** height of frame average per video
vii. **26** frame per second in each video on average

We have taken 20 different random categories from the dataset for our visulization purpose. For each selected action category we will select the first frame and display it with the label written on it. We can visualise 20 different videos from the data set before performing anything.

In order to evaluate the performance different approaches that we have implemented in the methodology section we will use different accuracy metrices. These metrics include accuracy, precision, recall and f-score. In our first approachwhich was really basic we just classify each frames of the video and among the predicted class for each we chose the majority one and label it as the out put of the video. So in the first method also we get fairly good accuracy of 87.26%. In the second method we got the accuracy ofaround 91 % where we have used 3D-CNN also known as slow fusion, also the f-score was 0.91. In our third approach which was using pose detection and LSTM we marked the accuracy at 92.51% and f-score at 0.92. In our final approach combining LSTM and CNN yield an accuracy of 92.19% with f-score of 0.92.

The deep learning algorithms are trained on online cloud resources because it were computational very expensiveand the saved model deployed on our web app to perform recognition on unseen data. To compare different algorithms and models there are different metrics available such as f-score, accuracy, precision and recall. Python package scikit-learn provide all these metrics to evaluate our model and predict which one is performing better.

The formula for calculation of these different metrics precision, recall, F-score and accuracy is given below:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F\_measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Like Xingjian Shi mentioned in his paper "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting" [13] To understand thebehaviour of his model he first compare

the ConvLSTM network with the FC-LSTM network using moving MNISTdataset. We run our model with different approaches and different kernel sized to study some of the cases in case of video recognition and sequence of data is important invideo recognition [21]. In the paper to verify the quality of the model on unseen problem Xingjian She build a new radar echo dataset and compared his model with ROVER algorithm. It was based on several nowcasting metrics for the more challenging nowcasting problems. Experiments results that was conducted on above two data sets leads to some of the findings. Two of them is mentioned below

While handling spatiotemporal correlations ConvLSTM is better than FC-LSTM.

While making state to state convolutional transition instead of all connected or full connected is required for capturing the motion patterns.

Better results has been produced by deeper model with fewer parameter.

ConvLSTM outperforms ROVER in case of precipitation nowcasting.

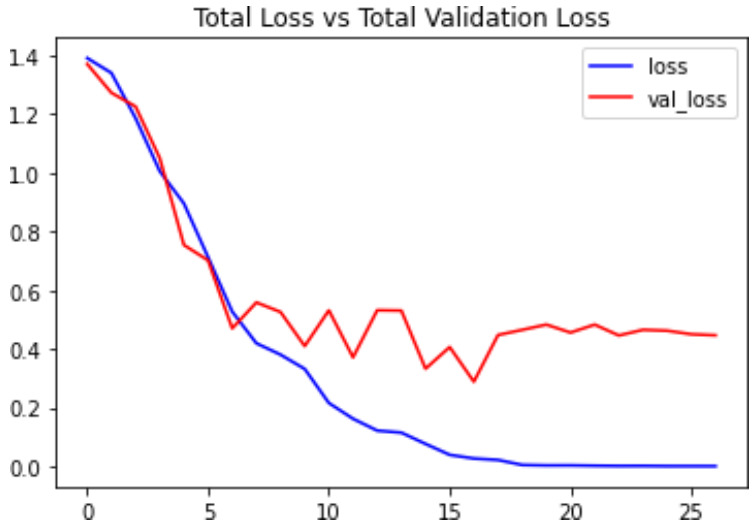
Below is the given summary of all the approaches that we have discussed above. These summary have scores of different metres used in each of the approaches. Clearly pose estimation with LSTM and combining LSTM with CNN gives the best accuracy.

4.1 Discussion

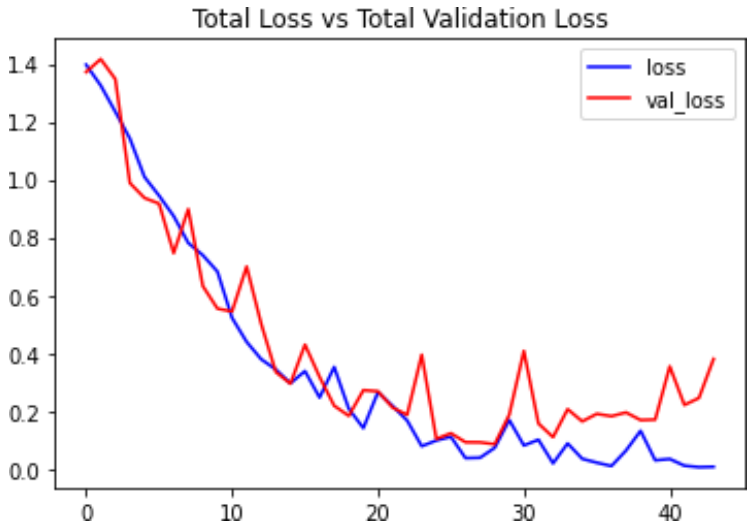
Combining CNN and LSTM outperforms all the other approaches by having accuracy of about 95%. By combining these two we extract the advantage of both and that explains the high accuracy. Spatial feature are extracted from the frames using convulational layers and then the extracted spatial feature is fed in LSTM layer. Thisenable each time steps for temporal sequence modeming.
This way a robust model is trained which learns onspatiotemporal features

The advantage with convLSTM is that it suitable for spatiotemporal data due to its inherent structure of Convolutional. By incorporating ConvLSTM into the encoding-forecasting structure, he build a model which can predict precipitation nowcasting. It is really intersting to know ConvLSTM can be applied to video-based action recognition. He mentioned that to add ConvLSTM on top of the spatial feature maps generated by a convolutional neural network and use the hidden states of ConvLSTM for the final classification could be one way of doing so. It is also evident from the accuracy metric graph of both the approach.

ConvLSTM Approach



LRCN Approach



5 Consclusion

In this paper, we have implemented 5 different approaches to perform video classification. We also learnedhow much important is temporal aspect of data to gain higher accuracy in video classification. We haveimplemented two convolutional neural network plus LSTMarchitectures in keras and tensor flow to perform Human Action Recognition on videos by using the temporal aswell as spatial information of the data. For pose estimation we have used OpenCV and Mediapipe through which we got some excellent results. As pose estimation have large use case Mediapipe provide immense area of application. One of the application is to detect the sign language through the videos.

Most of the current human activity recognition algorithmonly deals with single human subject. There is one limitation with our model which is it cannot work when there are multiple people present in the video. There shouldonly be a single person present in the video all the time doing some action so that our model can correctly recognise the action, this happens because the data we havechosen was in this manner, in all the videos in our data set only one person can be seen performing different activities.We can also use some other data set in which multiple people are present and train our model accordingly to perform activity recognition on multiple person videos. One quick way is to crop out each person and perform activity recognition separately on each of these persons butthis can be very very computational expensive. With the development in technologies better sensors will be available which will generate much more suitable dataset for these type of problems. And thus algorithms for multior group activity recognition will come up. Current application of activity recognition and pose estimate and vast use case but future applications may cover various aspects of a person's daily life from sports to healthy lifestyle and can absolutely change the way we interactwith our surroundings.

References

[1] William Taylor et al. "A Review of the State of the Art in Non-Contact Sensing for COVID-19". In: Sensors 20.19 (2020), p. 5665.

[2] Kia Dashtipour et al. "An Ensemble Based Classification Approach for Persian Sentiment Analysis". In: Progresses in Artificial Intelligence and Neural Systems. Springer, 2020, pp. 207–215.

[3] Mandar Gogate, Kia Dashtipour, and Amir Hussain. "Visual Speech In Real Noisy Environments (VISION): A Novel Benchmark Dataset and Deep Learning-based Baseline System". In: Proc. Interspeech 2020 (2020), pp. 4521–4525.

[4] Rami Ahmed et al. "Offline Arabic Handwriting Recognition Using Deep Machine Learning: A Review of Recent Advances". In: International Conference on

Brain Inspired Cognitive Systems. Springer. 2019, pp. 457– 468.

[5] Amir Hussain et al. “Artificial intelligence-enabled analysis of UK and US public attitudes on Facebook and Twitter towards COVID-19 vaccinations”. In: medRxiv (2020).

[6] Mandar Gogate, Ahsan Adeel, and Amir Hussain. “Deep learning driven multimodal fusion for automated deception detection”. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE. 2017, pp. 1–6.

[7] Metin Ozturk et al. “A novel deep learning driven, low- cost mobility prediction approach for 5G cellular networks: The case of the Control/Data Separation Architecture (CDSA)”. In: Neurocomputing 358 (2019), pp. 479–489.

[8] Mandar Gogate et al. “DNN driven speaker independent audio-visual mask estimation for speech separation”. In: arXiv preprint arXiv:1808.00060 (2018).

[9] Ahsan Adeel et al. “Lip-reading driven deep learning approach for speech enhancement”. In: IEEE Transactions on Emerging Topics in Computational Intelligence (2019).

[10] Mandar Gogate, Amir Hussain, and Kaizhu Huang. “Random Features and Random Neurons for BrainInspired Big Data Analytics”. In: 2019 International Conference on Data Mining Workshops (ICDMW). IEEE. 2019, pp. 522–529.

[11] Zheqi Yu et al. “Energy and performance trade-off optimization in heterogeneous computing via reinforcement learning”. In: Electronics 9.11 (2020), p. 1812.

[12] Mandar Gogate et al. “Av speech enhancement challenge using a real noisy corpus”. In: arXiv preprint arXiv:1910.00424 (2019).

[13] Kia Dashtipour et al. “Persent 2.0: Persian sentiment lexicon enriched with domain-specific words”. In: International Conference on Brain Inspired Cognitive Systems. Springer. 2019, pp. 497–509.

[14] William Taylor et al. “An intelligent non-invasive real- time human activity recognition system for

nextgeneration healthcare”. In: Sensors 20.9 (2020), p. 2653.

[15] Anis Koubaa et al. “Activity Monitoring of Islamic ˆ Prayer (Salat) Postures using Deep Learning”. In: 2020 6th Conference on Data Science and Machine Learning Applications (CDMA). IEEE. 2020, pp. 106–111.

[16] Ahsan Adeel et al. “A survey on the role of wireless sensor networks and IoT in disaster management”. In: Geological disaster monitoring based on sensor networks. Springer, 2019, pp. 57–66.

[17] Jaehyun Lee et al. “Automatic Classification of Squat Posture Using Inertial Sensors: Deep Learning Approach”. In: Sensors 20.2 (2020), p. 361.

[18] Fengling Jiang, Kia Dashtipour, and Amir Hussain. “A survey on deep learning for the routing layer of computer network”. In: 2019 UK/China Emerging Technologies (UCET). IEEE. 2019

[19] Fengling Jiang et al. “Robust visual saliency optimization based on bidirectional Markov chains”. In: Cognitive Computation (2020), pp. 1–12.

[20] Jinkun Han et al. “Lora-based smart IoT application for smart city: an example of human posture detection”. In: Wireless Communications and Mobile Computing 2020 (2020).

[21] Iulian Radu, Ethan Tu, and Bertrand Schneider. “Relationships Between Body Postures and Collaborative Learning States in an Augmented Reality Study”. In: International Conference on Artificial Intelligence in Education. Springer. 2020, pp. 257–262.

[22] Yash Agrawal, Yash Shah, and Abhishek Sharma. “Implementation of Machine Learning Technique for Identification of Yoga Poses”. In: 2020 IEEE 9th International Conference on Communication Systems and

Network Technologies (CSNT). IEEE. 2020, pp. 40–43.

[23] Muhammad Ali Imran Rami Ghannam and Qammer Abbasi. Engineering and Technology for Healthcare. Wiley-IEEE. ISBN 9781119644248, 2021.

[24] Intisar O Hussien, Kia Dashtipour, and Amir Hussain. “Comparison of sentiment analysis approaches using modern Arabic and Sudanese Dialect”. In: International Conference on Brain Inspired Cognitive Systems. Springer. 2018, pp. 615–624.