**M.Tech. Project Literature Survey**

**Name:** Rukhsar Sheikh

**M.Tech. Branch:** Artificial Intelligence and Robotics

**College Name:** College of Engineering Pune

**Project at Company:** R & DE Lab Dighi Pune

**Topic:** Human Activity Recognition Using Deep Learning Techniques

**College Guide**: Dr. S.M. Patil

**Company Guide:** Scientist Ajay sir

**Date: 15 /11/21**

**Contents:**

**Introduction**

**A] Research Papers Review**

1. 3D Convolutional Neural Networks for Human Action Recognition, Shuiwang Ji, Wei Xu, Ming Yang, Member, IEEE, and Kai Yu, Member, IEEE,2013.
2. Human Activity Recognition: A Review, Ong Chin Ann, Lau Bee Theng, IEEE International Conference on Control System, Computing and Engineering, 2014.
3. Human Action Recognition based on Convolutional Neural Networks with a Convolutional Auto-Encoder Chi Geng1, a, JianXin Song1, b 1 Nanjing University of Post and Telecommunications, Nanjing 210003, China, 2015.
4. Multi-channel Features Fitted 3D CNNs and LSTMs for Human Activity Recognition, Yang Qin, Lingfei Mo\*, Jing Ye, Zhening Du School of Instrument Science and Engineering Southeast University Nanjing, China, 2016.
5. Human Activity Recognition Using Robust Spatiotemporal Features and Convolutional Neural Network Md. Zia Uddin, Senior Member, IEEE, Weria Khaksar, and Jim Torresen, Senior Member, IEEE , 2017.
6. Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos J. Arunnehrua,∗ , G. Chamundeeswaria, S. Prasanna Bharath , International Conference on Robotics and Smart Manufacturing (RoSMa2018) ,2018.
7. Smart Phone Based Human Activity Recognition , Hongkai Chen, Sazia Mahfuz, Farhana Zulkernine School of Computing School of Computing Queen’s University Kingston, ON, Canada, 2019.
8. An Activity Recognition Framework for Overlapping Activities using Transfer Learning Muhammad Bilal Department of Computer Science, COMSATS University Islamabad, Attock Campus, Pakistan Mubashir Javaid Department of Computer Science, COMSATS University Islamabad, Attock Campus, Pakistan Muazzam Maqsood Department of Computer Science, COMSATS University Islamabad, Attock Campus, Pakistan Seungmin Rho Department of Software, Sejong University, Seoul, Korea Irfan Mehmood Department of Media Design and Technology, Faculty of Engineering & Informatics, University of Bradford, International Conference on Computational Science and Computational Intelligence (CSCI) ,2020.
9. Real-time Human Activity Recognition Using ResNet and 3D Convolutional Neural Networks ,Archana N. , Hareesh K. Dept. of Electronics and Communication Engineering Government College of Engineering Kannur Kannur, 2021.
10. An Efficient and Lightweight Deep Learning Model for Human Activity Recognition Using Smartphones Ankita 1 , Shalli Rani 1,\* , Himanshi Babbar 1 , Sonya Coleman 2,\*, Aman Singh 3 and Hani Moaiteq Aljahdali, 2021.

**B] Deep Learning Technique selection on the basis of accuracy.**

**C] Data Collection for making Dataset then labeling of dataset**

**D] Data Preprocessing (Coding)**

**E] Make a Model of selected deep learning technique (Coding)**

**F] Split the dataset into train and test set (Coding)**

**G] Train the model using training dataset (Coding)**

**H] Test the model using testing dataset (Coding)**

**I] Deployment of this model on real hardware system (Hardware experiment setup)**

**Date: 16 /11/21**

**Introduction**

The human assistance recognition system has been an essential part of today’s life of human. It recognizes the human’s presence or the current activity/action that relies upon that information that is being received from the various sensors (Cameras, accelerometer etc.). This system is used to detect the human’s motion daily activities.

A HAR system has major areas of applications, such as:

1. Interaction among humans and computers
2. Remote Monitoring
3. Military
4. Healthcare
5. Gaming
6. Sports
7. Security
8. Surveillance

This assistance system has other areas of wide applications, such as:

1. Monitoring Human’s Special Activities
2. Monitoring Sleep Disorders
3. Rehabilitation Activities Behavior Recognition
4. Fall Detection

There exist two physical activity recognition categories:

**(a)Vision-based:** The vision-based approach is limited to environments’ sensitivity of light and Detects at low range.

**(b) Non-vision based:** The non-vision approach is mainly used for human activity recognition that uses a different variety of sensors on wearable devices.

**Challenging task in HAR:**

1. The collection of huge data for HAR: Finding appropriate data and then preparing it for the further process makes use of costly sensors and issues related to privacy and security becomes difficult to use the public dataset.

2. The big challenge for HAR is the classification of time sequence task: A person’s movement can be predicted with the help of data acquired by the sensors and traditionally involves expertise in deep domain and methods to correctly design the features from the raw data so that a machine learning model can be fit into, from the signal processing method.

At present, deep learning methods, such as Convolutional Neural Networks (CNN) Recurrent Neural networks (RNN) and CNN-LSTM are very much capable of and have even achieved higher state-of-the-art results and have enabled the automatic learning of the features from the raw sensor data.

HAR is broadly classified into Sensor-based, Vision-based and multimodal categories:

**1. Sensor-based**— The first HAR approach contains a large number of sensor type technologies that can be worn on-body known as wearable sensors, ambient sensors, and, together, both will make hybrid sensors that help in measuring quantities of human body motion. Various opportunities can be provided by these sensor technologies which can improve the robustness of the data through which human activities can be detected and also provide the services based on sensed information from real-time environments, such as cyber-physical-social systems ; there is also a type of magnetic sensors when embedded in smartphone can track the positioning without any extra cost .

**2. Vision-based**—RGB video and depth cameras being used to obtain human actions.

**3. Multimodal**—Sensors data and visual data are being used to detect human activities. The evolvement of IoT is increasing rapidly. The major reason behind the evolvement is the compatibility of IoT with that of wearable sensors, network objects, and conventional networks . For example, the body sensor nodes are one of the most essential technologies of IoT by integrating both body sensor systems and networks of the wireless detector. There are various techniques used for information extraction. Deep Learning (DL) is an artificial intelligence (AI) function that uses structure-based algorithms functions, like a brain, to conclude and extract the important data from the given datasets. Deep Learning is an emerging field of research. The traditional machine learning approach requires feature extraction manually, but the deep learning approach performs automatic feature extraction. With the help of sensing devices, measured data were collected, results were analyzed, and a HAR system was developed.

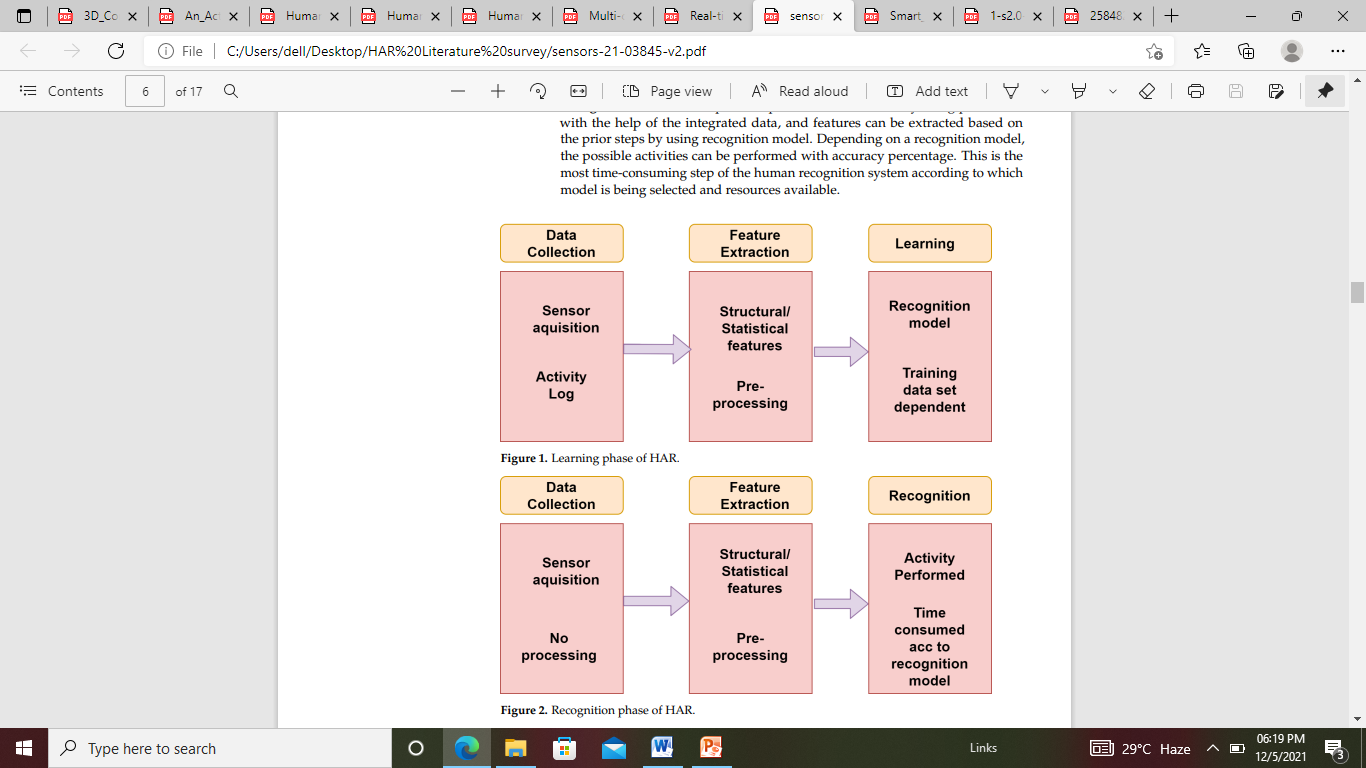


Fig: Learning Phase of HAR

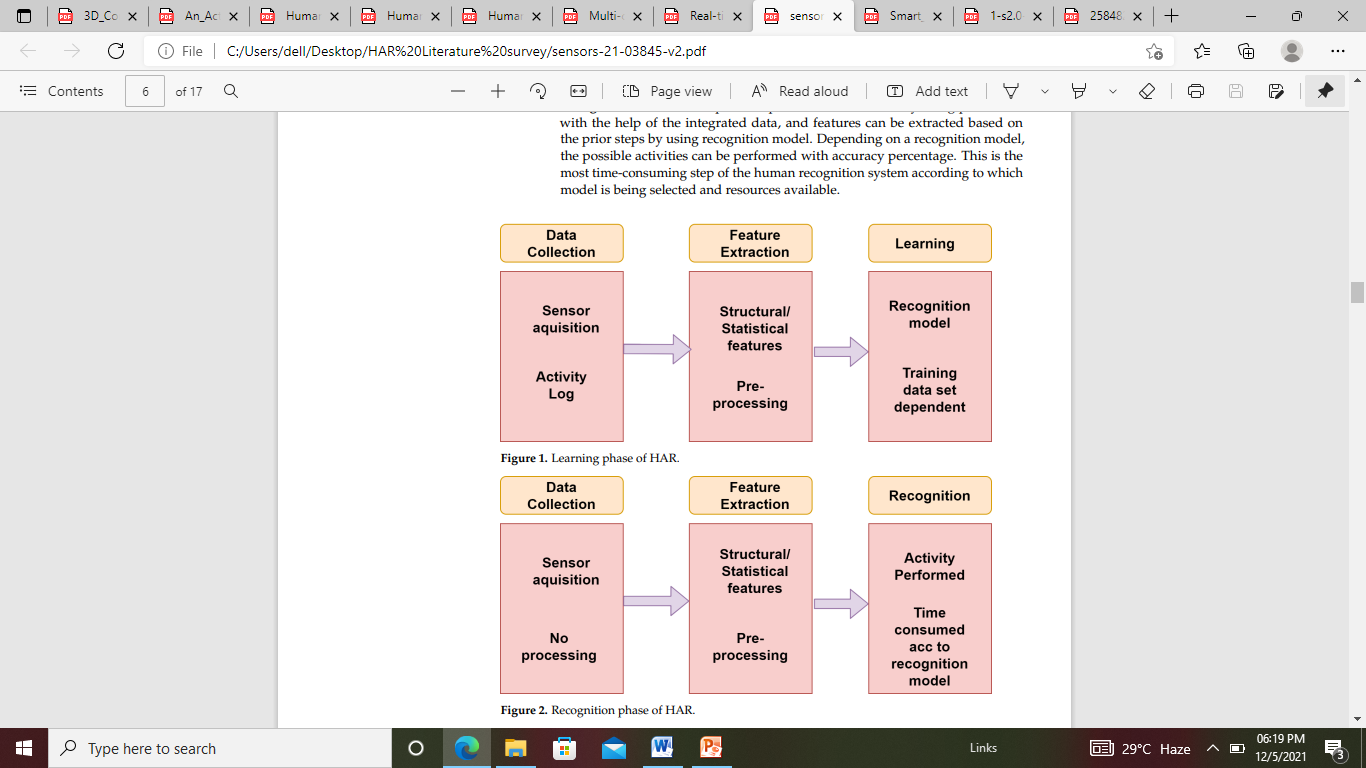


Fig: Recognition Phase of HAR

**Date:17/11/2**

**Research Papers Review**

1. **3D Convolutional Neural Networks for Human Action Recognition,**

Shuiwang Ji, Wei Xu, Ming Yang, Member, IEEE, and Kai Yu, Member, IEEE,2013.

**Topic:** Human Activity Recognition using 3D CNN

**Dataset:** TREC Video Retrieval Evaluation (TRECVID) dataset (surveillance video data recorded at London Gatwick Airport)

**Deep Learning Model Used:** 3D CNN architecture

In this paper they develop a 3D CNN architecture that generates multiple channels of information from adjacent video frames and performs convolution and subsampling separately in each channel. The final feature representation is obtained by combining information from all channels. They propose to apply the 3D convolution operation to extract spatial and temporal features from video data for action recognition. These 3D feature extractors operate in both the spatial and the temporal dimensions, thus capturing motion information in video streams.

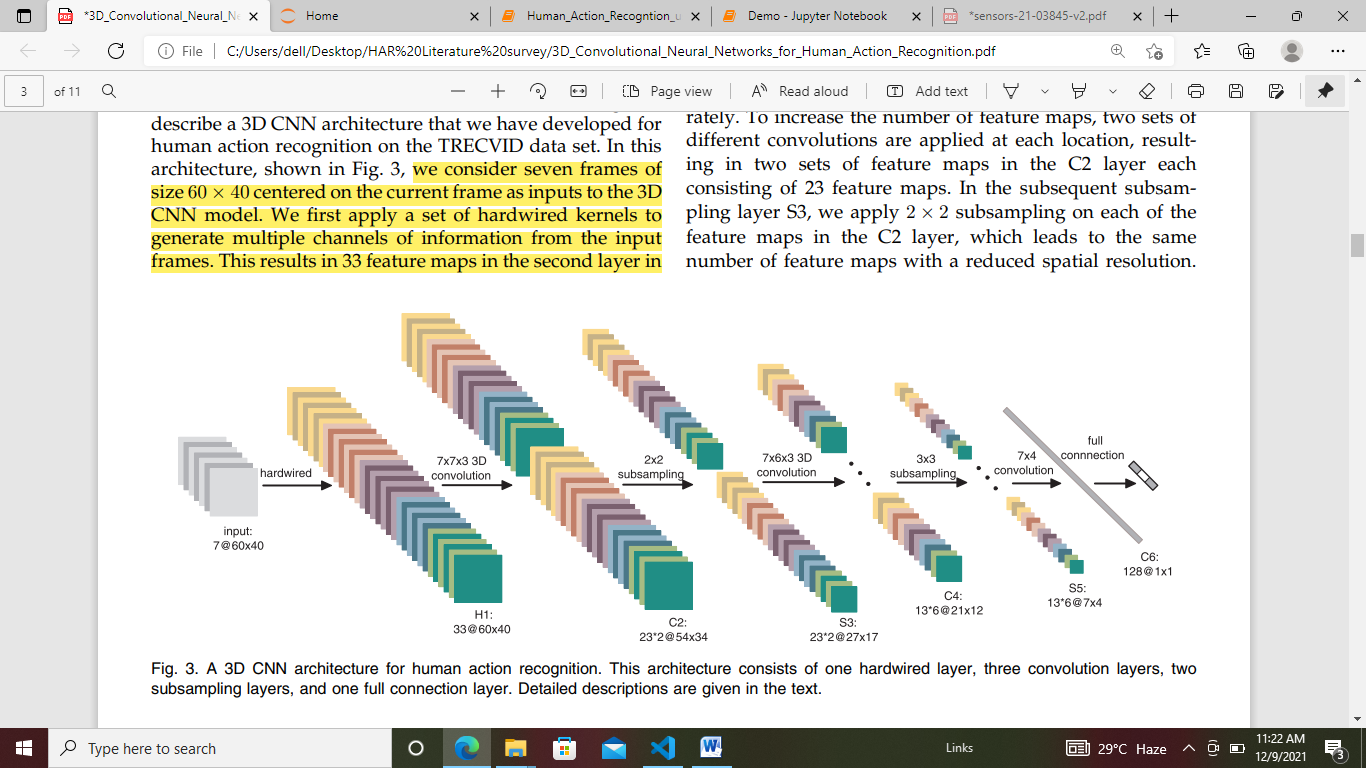


Fig. 3. A 3D CNN architecture for human action recognition.

This architecture consists of one hardwired layer, three convolution layers, two subsampling layers, and one full connection layer.

**Date: 18/11/21**

**3D CNN Architecture:**

Here they consider seven frames of size 60 x40 centered on the current frame as inputs to the 3D CNN model. They first apply a set of hardwired kernels to generate multiple channels of information from the input frames. This results in 33 feature maps in the second layer in five different channels denoted by gray, gradient-x, gradienty, optflow-x, and optflow-y. The gray channel contains the gray pixel values of the seven input frames. The feature maps in the gradient-x and gradient-y channels are obtained by computing gradients along the horizontal and vertical directions, respectively, on each of the seven input frames, and the optflow-x and optflow-y channels contain the optical flow fields along the horizontal and vertical directions, respectively, computed from adjacent input frames.

**1. Model Regularization:** The inputs to 3D CNN models are limited to a small number of contiguous video frames due to the increased number of trainable parameters as the size of input window increases

**2. Model Combination:** In the prediction phase, the input is given to each model and the outputs of these models are then combined. Experimental results demonstrate that this model combination scheme is very effective in boosting the performance of 3D CNN models on action recognition tasks.

**3. Model Implementation:** The 3D CNN models are implemented in C++ as part of NEC’s human action recognition system. All the model parameters are randomly initialized and are trained using the stochastic diagonal Levenberg-Marquardt method. In this method, a learning rate is computed for each parameter using the diagonal terms of an estimate of the Gauss-Newton approximation to the Hessian matrix on 1,000 randomly sampled training instances.

**Conclusion:** the 3D CNN model outperforms compared methods on the TRECVID data.

**Drawback:** The developed 3D CNN model was trained using a supervised algorithm in this paper, and it requires a large number of labeled samples. Prior studies show that the number of labeled samples can be significantly reduced when such a model is pretrained using unsupervised algorithms. We will explore the unsupervised training of 3D CNN models in the future.

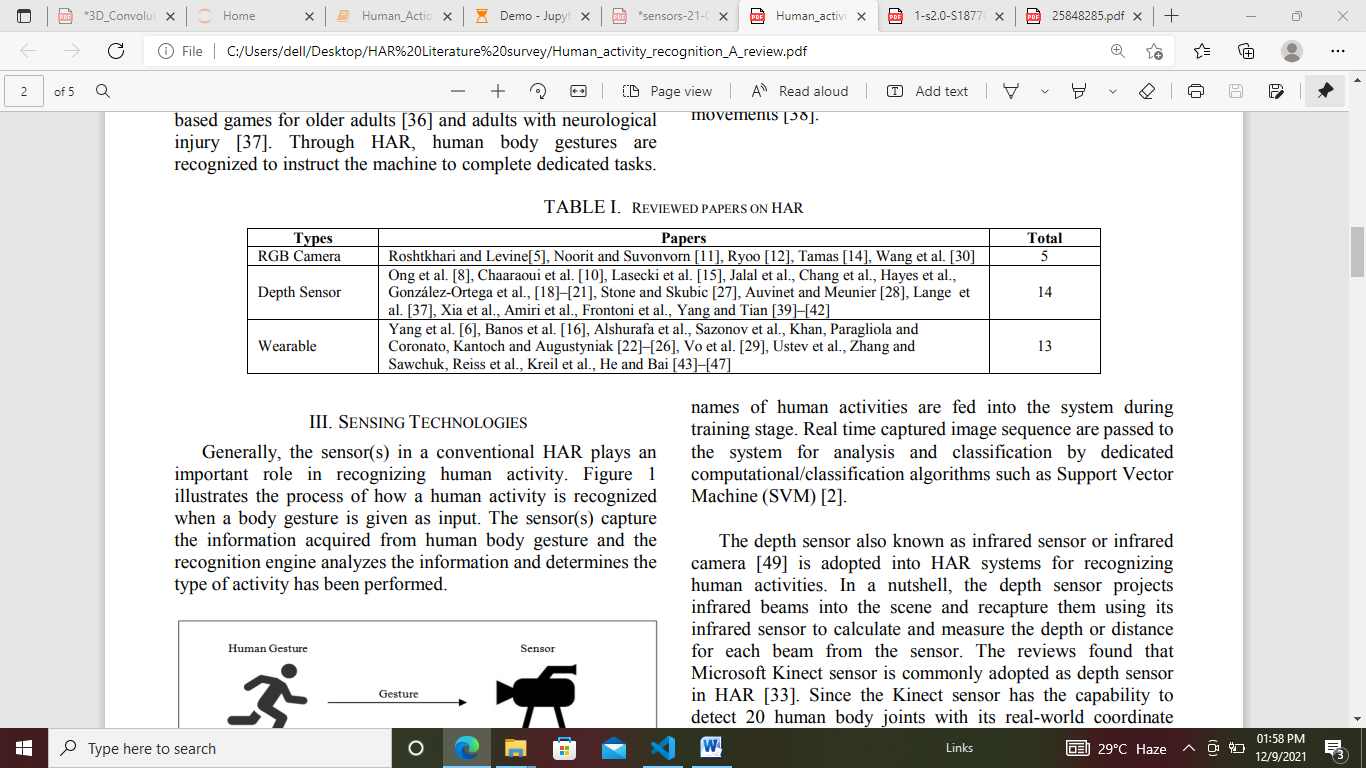
**Date: 19/11/21**

1. **Human Activity Recognition: A Review**

Ong Chin Ann, Lau Bee Theng, IEEE International Conference on Control System, Computing and Engineering, 2014.

**Topic: Review of 32 HAR research paper from 2011-2014**

We reviewed 32 papers published recently (from 2011 to 2014) on different sensing technologies used in HAR. Most of the conventional HAR systems using this sensing technology are built with two major components which is the feature extraction and classification. These technologies are classified as RGB camera-based, depth sensor-based and wearable-based as shown in Table I.



**RGB camera** is simple but with low efficiency. A RGB camera is usually attached to the environment and the HAR system will process image sequences captured with the camera. Besides, most of the RGB-HAR systems are considered as supervised system where trainings are usually needed prior to actual use. Image sequences and names of human activities are fed into the system during training stage.

**Depth sensor** also known as infrared sensor or infrared camera. In a nutshell, the depth sensor projects infrared beams into the scene and recapture them using its infrared sensor to calculate and measure the depth or distance for each beam from the sensor. The reviews found that Microsoft Kinect sensor is commonly adopted as depth sensor in HAR. Since the Kinect sensor has the capability to detect 20 human body joints with its real-world coordinate, many researchers utilized the coordinates for human activity classification.

**Wearable-based** requires single or multiple sensors to be attached to the human body. Most commonly used sensor includes 3D-axial accelerometer, magnetometer, gyroscope and RFID tag. With the advancement of current smart phone technologies, many works uses mobile phone as sensing devices because most smart phones are equipped with accelerometer,

**Date: 22/11/21**

Magnetometer and gyroscope. A physical human activity can be identify easily through analyzing the data generated from various wearable sensing after being process and determine by classification algorithm.

The review outcome indicates both depth sensor and wearable sensor technologies are gaining more popularity in HAR research recently. On the other hand, RGB camera has obtained less emphasis in HAR research, most probably due its imitation in capturing the scene and human motions in 3D space.

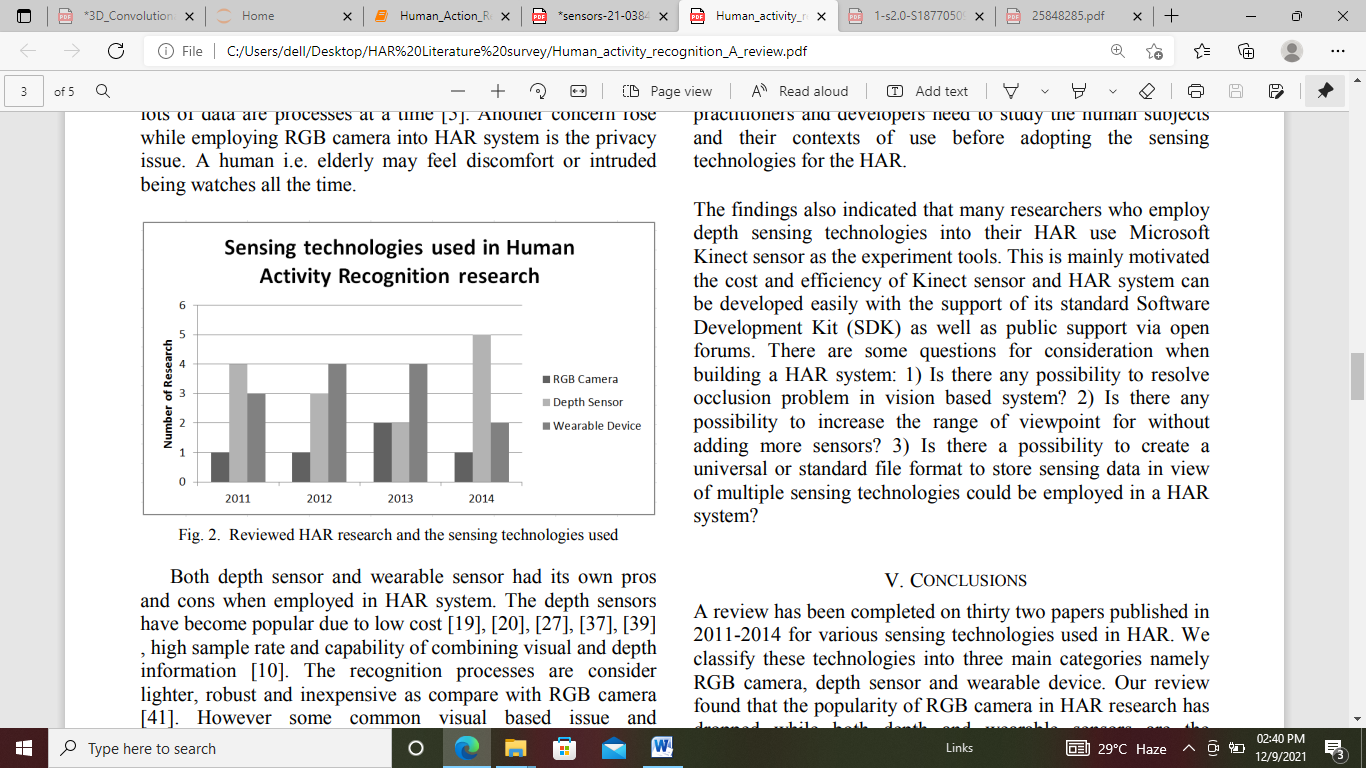


Fig. Reviewed HAR research and the sensing technologies used

**Conclusion:** Our review found that the popularity of RGB camera in HAR research has dropped while both depth and wearable sensors are the substitutes. On the other hand, the use of Kinect sensor (depth sensor) into HAR system is promising. This could be a sign of the rise of Kinect as a popular sensing tool in HAR system.

**Date: 23/11/21**

1. **Human Action Recognition based on Convolutional Neural Networks with a Convolutional Auto-Encoder**

Chi Geng1, a, JianXin Song1, b 1 Nanjing University of Post and Telecommunications, Nanjing 210003, China, 2015.

**Topic:** Human Action Recognition based on Convolutional Neural Networks.

**Dataset:** KTH dataset.

**Proposed Model:** CNN Model and Pretrained CNN Model**.**

* Proposed model in HAR that can act directly on the raw inputs.
* An efficient pre-training strategy has been introduced to reduce the high computational cost of kernel training to enable improved real-world applications.
* CNN has invariance for a particular pose, illumination, and disorderly environmental change.

**A generic description of human action recognition from image sequence consists of two steps**:

1) Extract complex handcrafted features from raw input video frames

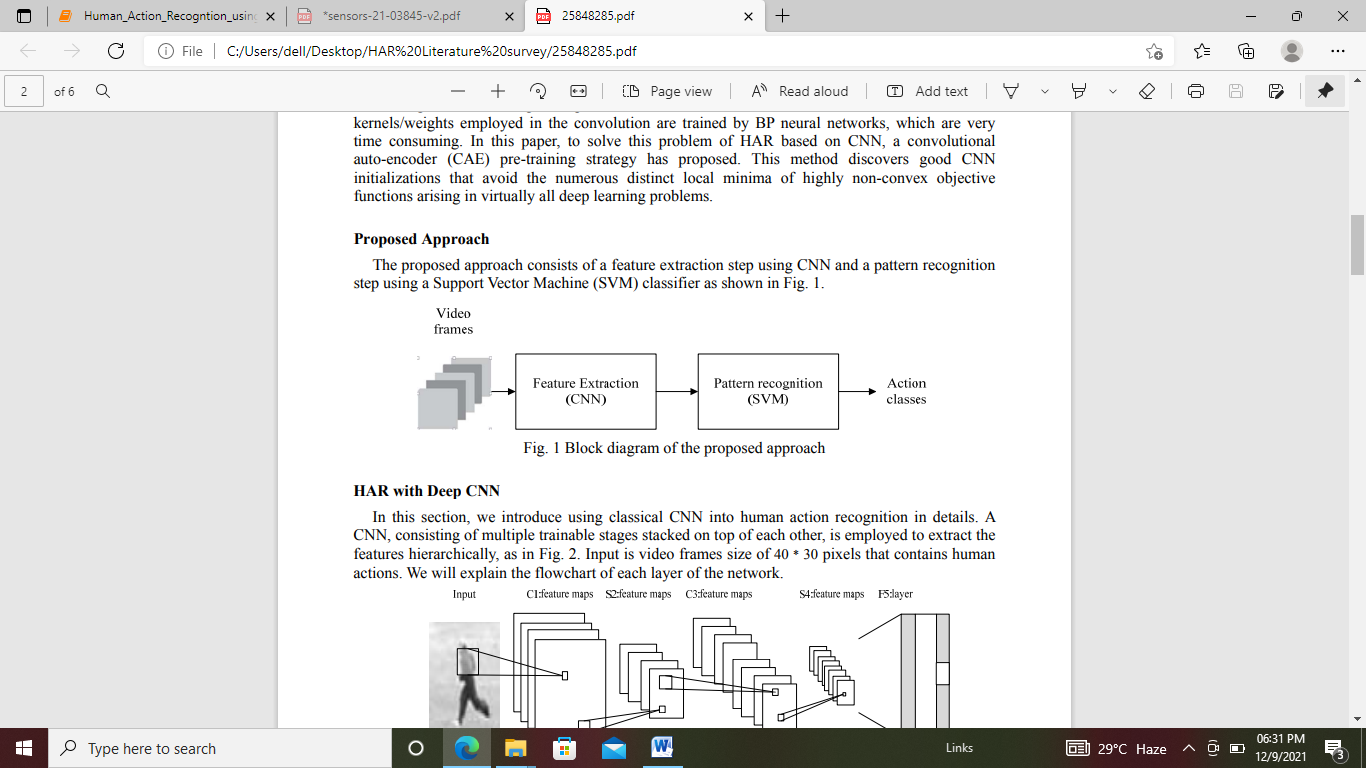
2) Build a classifier based on these features some of the commonly used features for human action recognition are histogram of oriented gradient (hog)[1], histogram of optical flow (hof) , motion interchange patters (mip), space-time interest points (stip), action bank features and dense trajectories.

This method still has a weakness that the kernels/weights employed in the convolution are trained by BP neural networks, which are very time consuming.

In this paper, to solve this problem of HAR based on CNN, a convolutional auto-encoder (CAE) pre-training strategy has proposed. This method discovers good CNN initializations that avoid the numerous distinct local minima of highly non-convex objective functions arising in virtually all deep learning problems

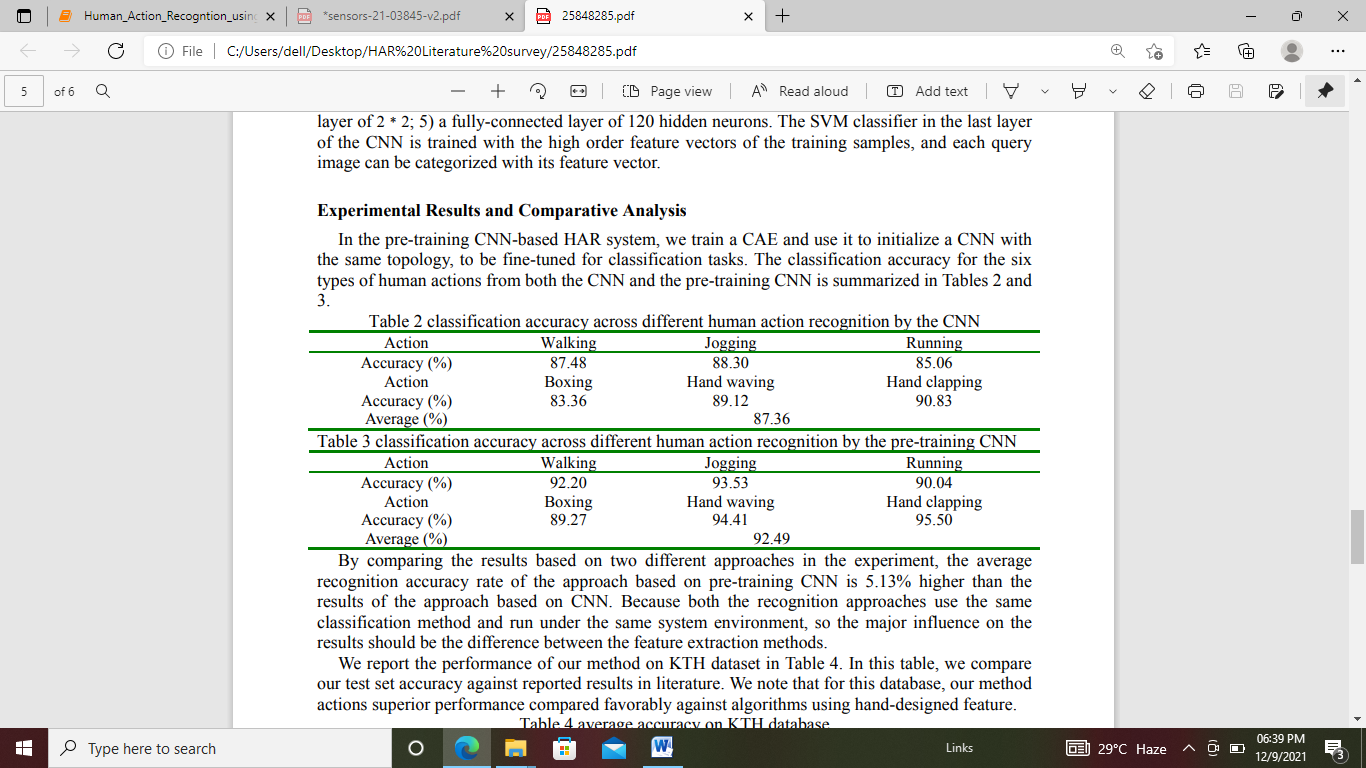
Proposed Approach The proposed approach consists of a feature extraction step using CNN and a pattern recognition step using a Support Vector Machine (SVM) classifier as shown in Fig. 1.

**Date: 24/11/21**



In CNN-based HAR model, parameters of each layer in the CNN The network has 6 hidden layers: 1) convolutional layer with 6 5﹡5 filters per input channel; 2) max-pooling layer of 2﹡2; 3) convolutional layer with 16 5﹡5 filters per map; 4) max-pooling 936 layer of 2﹡2; 5) a fully-connected layer of 120 hidden neurons. The SVM classifier in the last layer of the CNN is trained with the high order feature vectors of the training samples, and each query image can be categorized with its feature vector.

In the pre-training CNN-based HAR system, we train a CAE and use it to initialize a CNN with the same topology, to be fine-tuned for classification tasks. The classification accuracy for the six types of human actions from both the CNN and the pre-training CNN is summarized in Tables 2 and 3.



By comparing the results based on two different approaches in the experiment, the average recognition accuracy rate of the approach based on pre-training CNN is 5.13% higher than the results of the approach based on CNN. Because both the recognition approaches use the same classification method and run under the same system environment, so the major influence on the results should be the difference between the feature extraction methods.

**Date: 25/11/21**

1. **Multi-channel Features Fitted 3D CNNs and LSTMs for Human Activity Recognition,**

Yang Qin, Lingfei Mo\*, Jing Ye, Zhening Du School of Instrument Science and Engineering Southeast University Nanjing, China, 2016**.**

**Topic:** 3D CNNs and LSTMs fusion model in human activity recognition

**Dataset:** KTH dataset

**Reveiw**

**Multi-channel features**

**A. Motion Optical Flow Vector**

Motion optical flow vector is used to describe motion coherence. The motion optical flow vector is an independently feature without any knowledge of human body and shape information.

**B. Grey Scale**

Gray scale is not sensitive to noise because only substantial original image information needs to be preserved, which means it can compensate for other features which are sensitive to noise

**C. Body Edge Image**

The time-varying characteristic of body shape is used to classify activities, because the model has Long-Short Term Memory neural network, which can extract information of time series efficiently.

**Model structure**

**A. System Work flow**

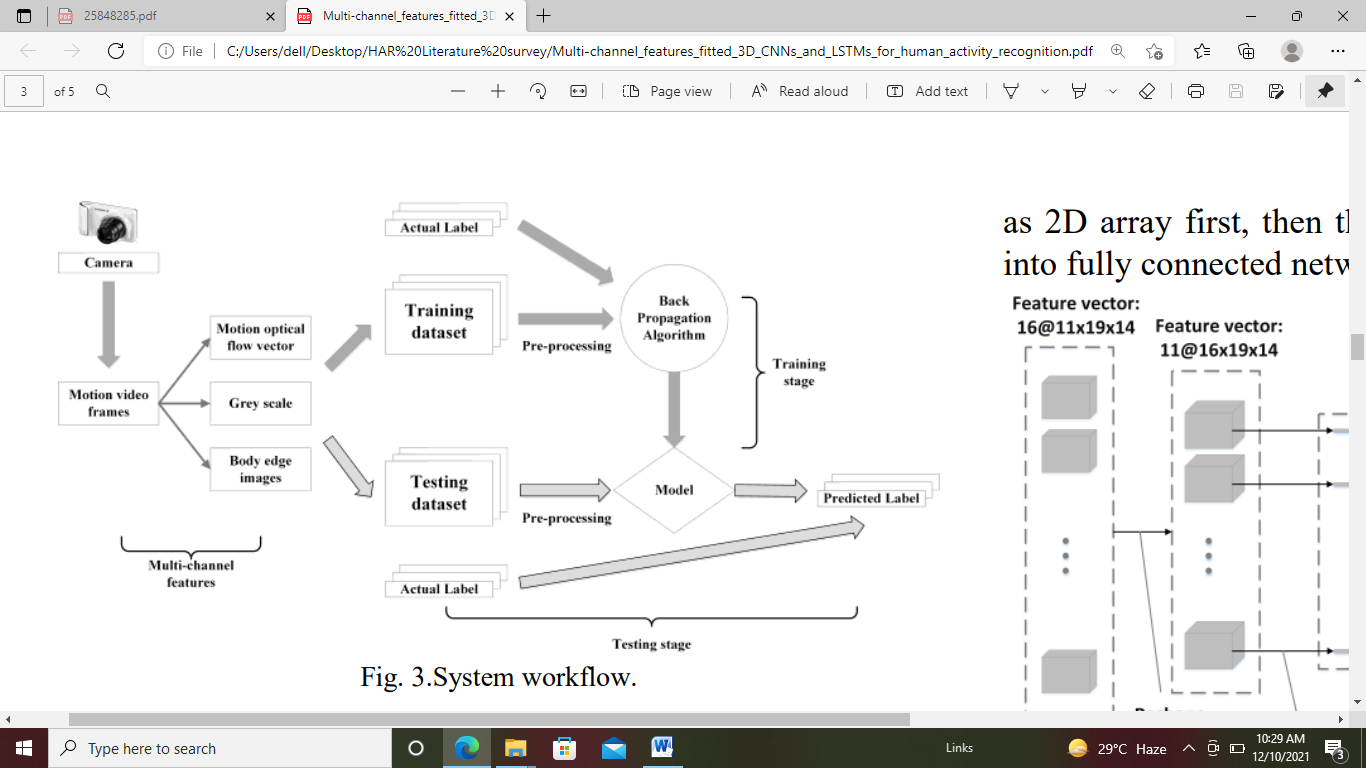


Fig.System workflow

**Date:26/11/21**

**B. Deep Neural Network**

1) 3D Convolutional Neural Network.

2) Reshape Layers

3) Long-short Term Memory Neural Network

**Network Architecture**

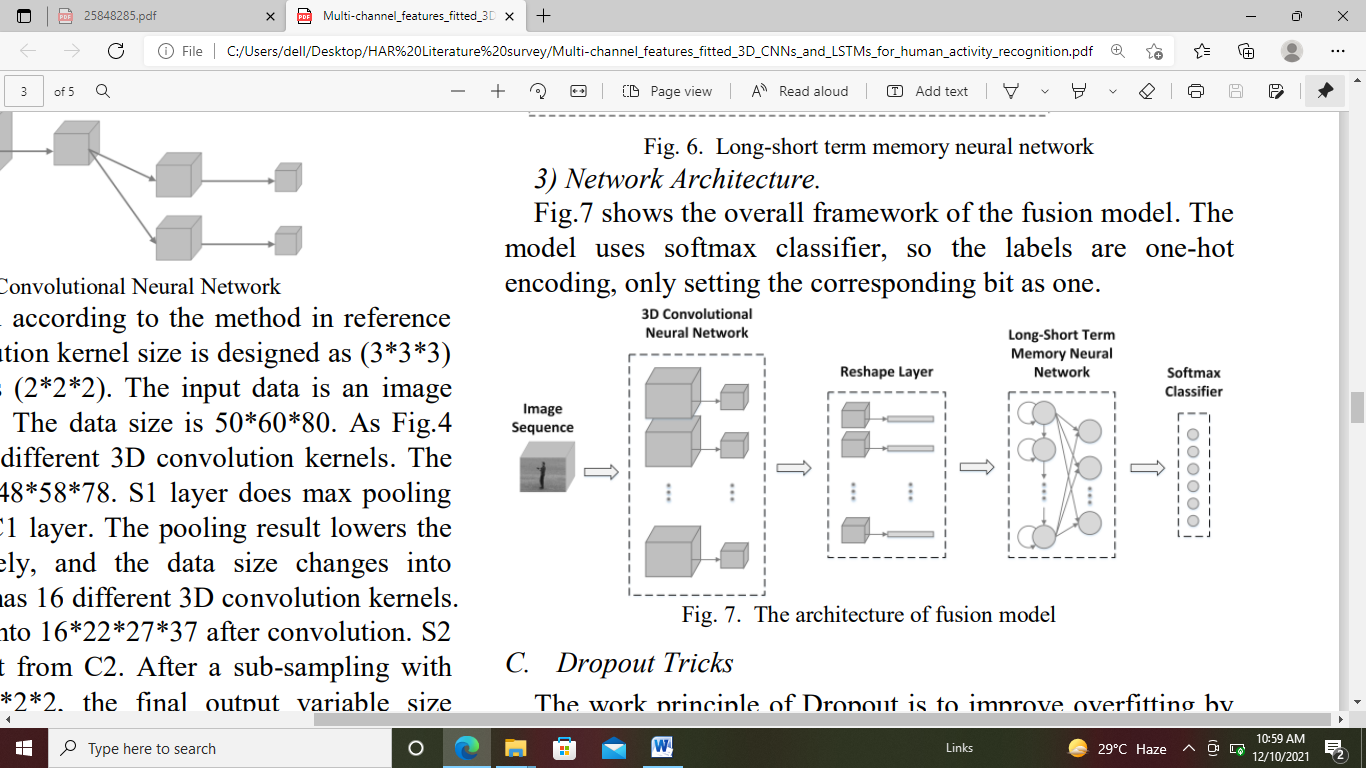


Fig.The architecture of fusion model

**C. Dropout Tricks**: The work principle of Dropout is to improve over fitting by preventing co-adaptation of feature detectors.

**EXPERIMENT AND RESULTS**

**A. KTH Description**

**B. Data Processing**

1) Gaussian Filtering

2) Standardizing Input Data

3) Shuffling Samples

4) K-fold Cross Validation

**C. Framework**

The paper uses python to build the model with Tensor Flow. The environment of experiment contains NVIDIA GTX 960 1024 CUDA cores / Intel i3 2.26GHz / 8GB DDR3 Memory / Ubuntu 14.01 x64. D.

**Results and Analysis**

The average recognition rate of the model based on RGB information is 90.8%, while using multichannel feature can increase the average recognition rate to 94.3%.

**Date: 29/11/21**

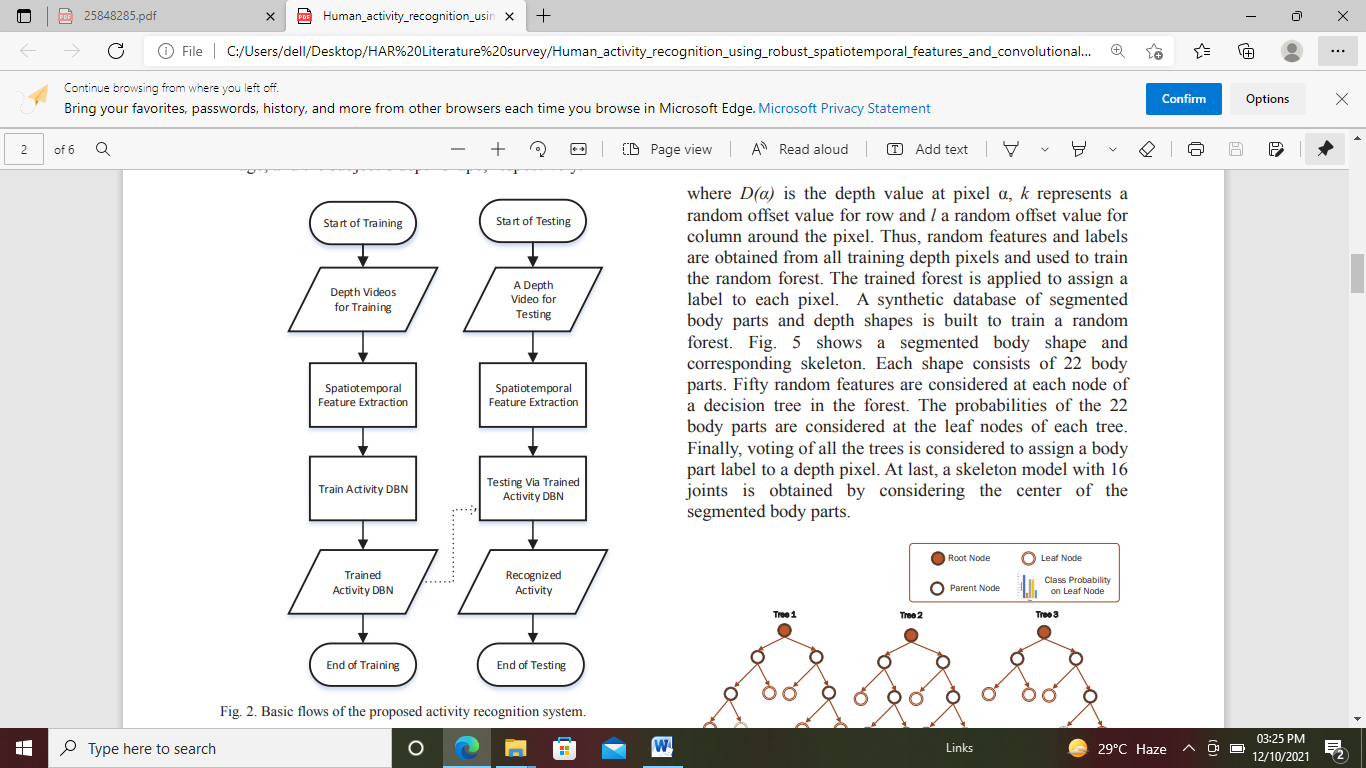
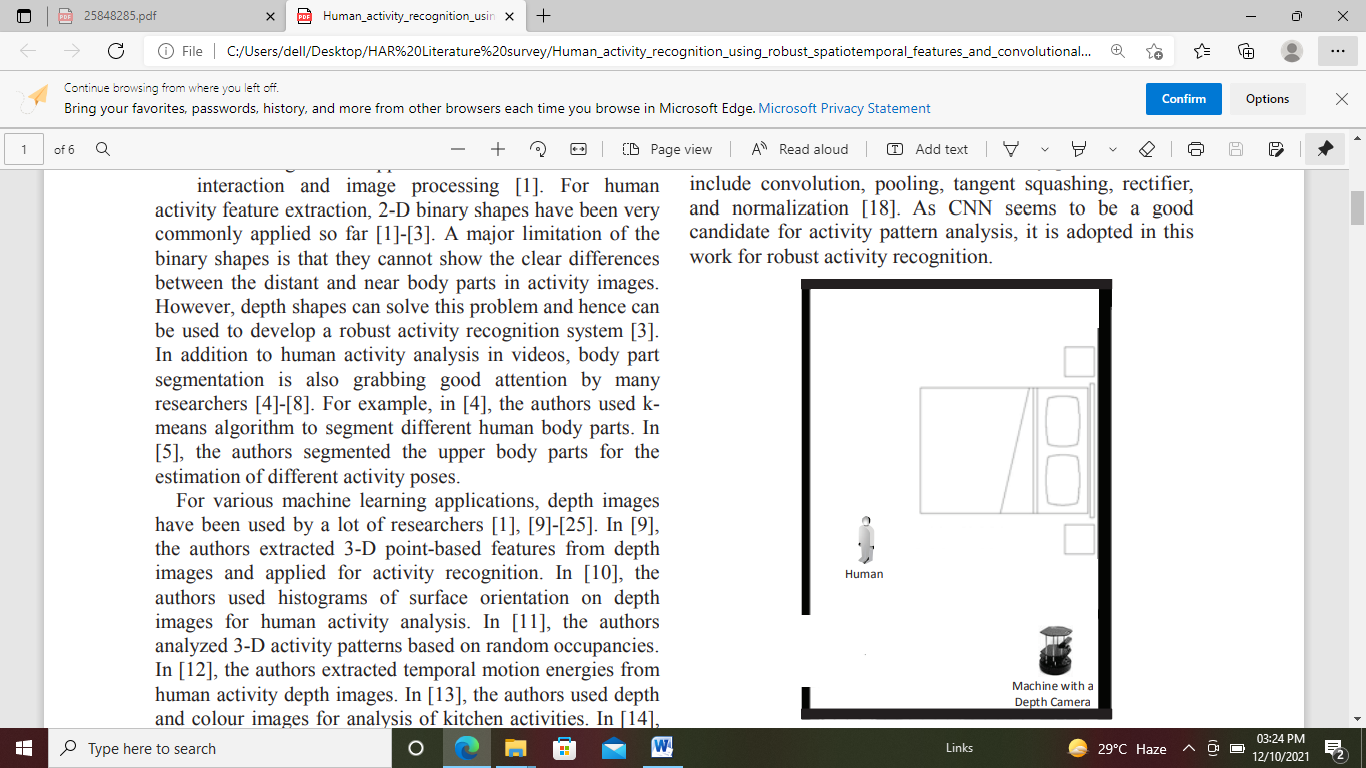
**5. Human Activity Recognition Using Robust Spatiotemporal Features and Convolutional Neural Network**

Md. ZiaUddin, Senior Member, IEEE, Weria Khaksar, and Jim Torresen, Senior Member, IEEE,2017.

**Topic:** A novel approach has been proposed for depth camera-based human activity recognition utilizing robust spatiotemporal features and 1-D CNN.

**Dataset:** MSRC-12 Gesture Dataset

**Methodology:** The proposed human activity recognition system consists of several steps including depth video processing, feature generation, and modeling activity via CNN.



(1) (2)

Fig.1. A schematic room setup for depth camera-based activity recognition.

Fig.2. Basic flows of the proposed activity recognition system

**A. Body Part Segmentation**: A random forest used in this work for the segmentation process of different body parts. The first consists of some decision trees. Each tree in the forest consists of nodes and leaves.

**B. Feature Generation**: Once we obtain the segmented body shape, the 3-D centroid of each body part is combined to represent the 3-D skeleton model.

**Date: 30/11/21**

**1. Spatial Features:** From a depth image, the first spatial features are represented by the angles from one joint to another. As there are 16 joints in a skeleton, there are 15\*15=225 possible joint pairs.

**2. Temporal Features:** For an activity frame, the temporal features are extracted considering the magnitude and direction of the 16 body joints in next frame.

**C. Convolutional Neural Network** for Activity Modeling Convolutional Neural Network (CNN) is mostly used for image-based deep learning applications [27]. Compared to other deep learning structures, CNN often demonstrates better recognition performance in machine vision applications due to its ability to extract and learn imagebased features. Besides, CNN has the advantage of using a small amount of bias and weight values than other deep learning methods.

**Experiments and results**

A database of six activities was built to check different human activity recognition approaches. The activities were left leg moving, right leg moving, both hand waving, right hand waving, sitting-down, and standing-up. One hundred clips from each activity were collected for the training purpose. Finally, 100 clips were used to test each activity.

The spatiotemporal features were combined with HMM which achieved 91.33% mean recognition rate. Later on, the proposed approach (i.e., the spatiotemporal features with CNN) was tried that achieved the highest recognition rate (i.e., 98.17%), showing its superiority over all other approaches.

**Date: 01/12/21**

**6. Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos**

J. Arunnehrua,∗ , G. Chamundeeswaria, S. Prasanna Bharath , International Conference on Robotics and Smart Manufacturing (RoSMa2018) ,2018.

**Topic**: 3D - Convolutional Neural Networks (3D-CNN) with 3D motion cuboid for action detection and recognizing in videos. The proposed method is compared with the existing methods in terms of accuracy.

**Dataset:** The experiments are conducted on benchmark KTH and Weizmann dataset.

**3D - Motion Cuboid**: Frame difference approach is immensely adaptive to identify the motion scene similar to moving objects in the dynamic environment. The absolute temporal difference frame is attained by subtracting the earlier frame t with current frame t+ 1 on a pixel by pixel basis. Fig. 1 shows the continuous frames of the ’running’ action from KTH dataset. The resulting difference frame stack is called as 3D Motion cuboid.

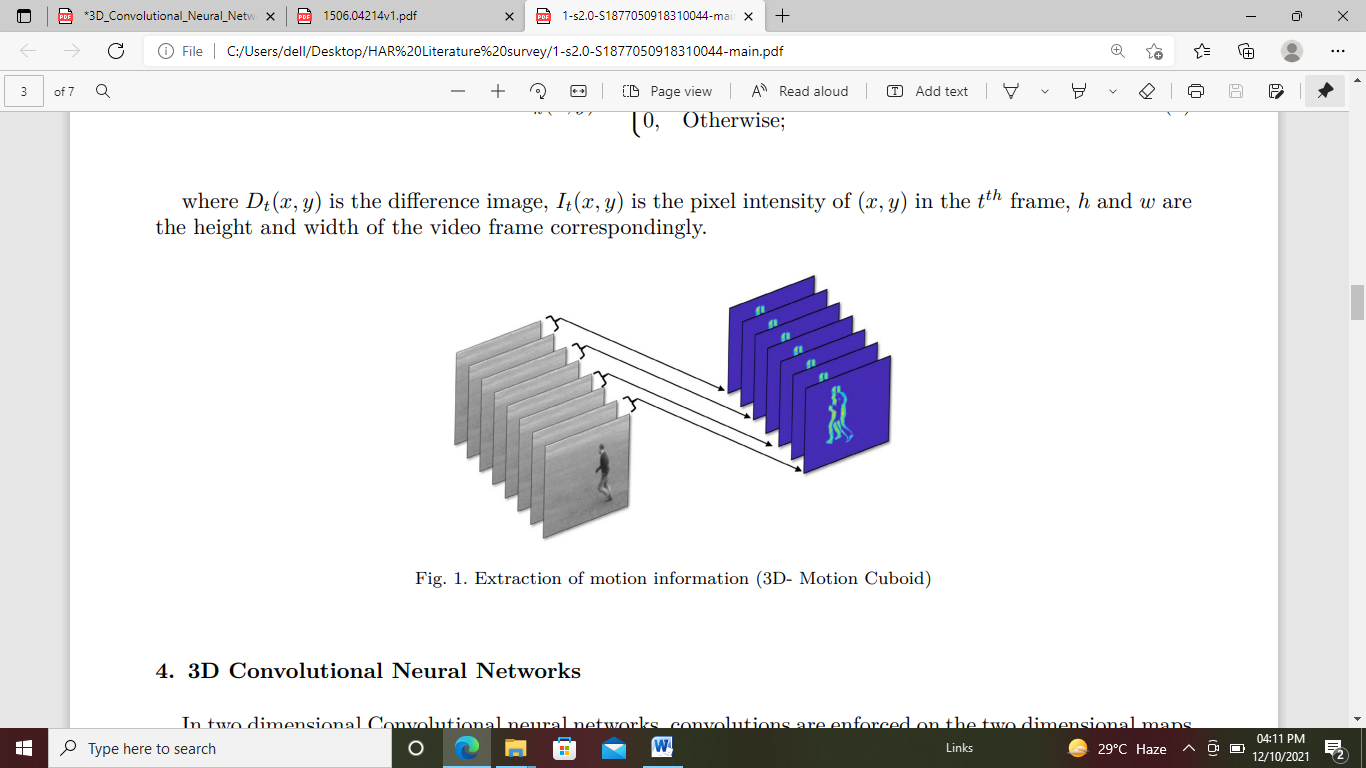


Fig.1. Extraction of motion information (3D- Motion Cuboid)

**3D Convolutional Neural Networks:** They introduce to enumerate three dimensional convolutions in the succeeding stages of CNNs to gauge the features from both the temporal and spatial dimensions. The 3D convolution is obtained by convolving a three dimension kernel to the cube obtained by assembling more than one spatial temporal patches arranged in a contiguous manner. The feature maps present in the convolution layer is linked with the multiple frames arranged contiguously in the previous layer in order to capture the motion related information. It is noted that 3D convolution kernel can select only one type of feature from the patch cuboid, provided the kernel weights are duplicated across the patch cube.

**Date: 02/12/21**

**Experimental Setup**: The analysis is conducted in Python 3.5 with OpenCV 3.1 and Tensorflow in Windows 10 OS on PC with Intel i7 processor with RAM 8 GB.

**Experimental Results:** On KTH dataset where correct predictions are most of the actions classes like running, jogging, single hand waving, clapping and boxing are almost predicted well with more than 94%. An average recognition rate of 3D-CNN classifier on KTH dataset is 94.9%. From the confusion matrix, the walking action is misclassified as running due to its similar pattern.

**Experimental Results:** On Weizmann dataset most of the action classes like walk, side, skip, jump, pjump, jack, bend and wave with one hand are almost predicted well. An average recognition rate of 3D-CNN classifier on Weizmann dataset is 97.2%. From this, wave with both hands action is misclassified as walk and wave with one hand respectively.

**Comparative Study**: The results obtained by the 3D-CNN with motion cuboid are compared quantitatively with state-of-theart results with KTH and Weizmann dataset to measure the effectiveness of the proposed action recognition system and comparison is presented Based on the comparison, it is seen that the proposed method shows best results on KTH and Weizmann action dataset.

**Date: 03/12/21**

**7. Smart Phone Based Human Activity Recognition**

Hongkai Chen, Sazia Mahfuz, Farhana Zulkernine School of Computing School of Computing Queen’s University Kingston, ON, Canada, 2019.

**Topic:** 3D tensor data structures from 2D time series data obtained from multiple sensors on a smart phone, and a new Convolutional Neural Network (CNN) model, which uses the tensor data and performs automatic feature extraction and classification for HAR.

**Dataset:** MobiAct v2.0 HAR dataset from Biomedical Informatics and eHealth Laboratory (BMI lab).

**Input:** Sensor data (e.g. Accelerometer in smart phone)

**Model:** Convolutional Neural Network (CNN)

Our tensor and CNN models achieved an overall better performance in classifying 11 Activity of Daily Living (ADL) than the state-of-the-art approaches and 15% higher sensitivity compared to previous approaches

**Classifier:** K-Nearest Neighbors (kNB)( best classifier with best accuracy), J48 decision tree, Logistic regression and Multilayer perceptron.

The results illustrated that kNB had a comparatively better performance with 99.88% accuracy in classifying six basic ADLs (Activity of Daily Living).

**There are still ongoing challenges that need to be addressed:**

(1) The selection of sensor data

(2) Extraction of relevant feature set

(3) Choice of the classifier

(4) General applicability of pre-trained classification models

(5) Power limitations of mobile devices.

Three variances of ANNs were implemented for classification: a multi-layer perceptron with backpropagation learning using the Neuroph framework, a feed forward ANN with backpropagation learning using the Encog framework, and a deep ANN or DNN using DeepLearning4j. The DNN with normalized data always presented a higher accuracy (80%) than the other classifiers. However, the use of a new dataset with only ANNs makes it difficult to compare this work with other research.

**Date: 06/12/21**

The authors performed feature selection using the WEKA toolkit and examined several machine learning methods including Naïve Bayes, LogitBoost, Random Forest and Support Vector Machine (SVM) to select the most reliable classifier. Random Forest had the highest average accuracy (92.29%) when both accelerometer data and gyroscope data were used.

A novel architecture to treat each axis of the accelerometer as one channel of the RGB image. The CNN was applied separately to each data channel to construct a feature set with strong local dependencies. The max-pooling layer proved to be capable of preserving scale invariance, which is crucial for real-world implementations. The CNN partial weight sharing model had the highest classification accuracy of 88.19%, 76.83%, 96.88% on the publicly available Skoda, Opportunity, and Actitracker datasets respectively, which is 4.41%, 1.2%, 9.02% higher than the PCA-ECDF (Principal Component Analysis - Empirical Cumulative Distribution Function) algorithm.

The CNN had the highest accuracy of 93.7% for the PAMAP2 dataset which has comparable data types as the MobiAct dataset. The LSTM and b-LSTM showed the most promising results of 76% and 92.7% accuracy for the DG and OPP datasets.

**Implementation**

**A. Dataset**

**B. Environment:** the implementation environment used in our research is discussed here. We used the Ryzen 5 2600k as our CPU; DDR4 16G as the RAM; RTX 2070 8G as the GPU; Windows 10 as the operating system; Anaconda as the software platform; and Python 3 as the programming language. Based on the comparative study of different deep learning software tools by TensorFlow due to its wide spread use, relatively low complexity and training time, fast testing speed, support for GPU and scalability. Anaconda was used on Windows 10 operating system to set up the above tools and Anaconda Accelerate was used to get more options to optimize the environment to use Intel CPUs and NVIDIA GPUs.

**C. Data processing:** To process the raw multi-sensor time series data from the MobiAct v2.0 dataset, we first applied a down sampling process to reduce the sampling frequency and then a linear interpolation to construct the new measure points in the raw data. The original sampling frequency of MobiAct v2.0 dataset was 200Hz, which was too high for HAR research. According to the survey on HAR using mobile , 20Hz is sufficient, and a common and suitable

**Date: 07/12/21**

Sampling frequency for a HAR study. For this study, we used 25Hz, 50Hz, and 100Hz to test different approaches based on the structure of our feature map.

**D. Definition of Tensor Models for CNN**

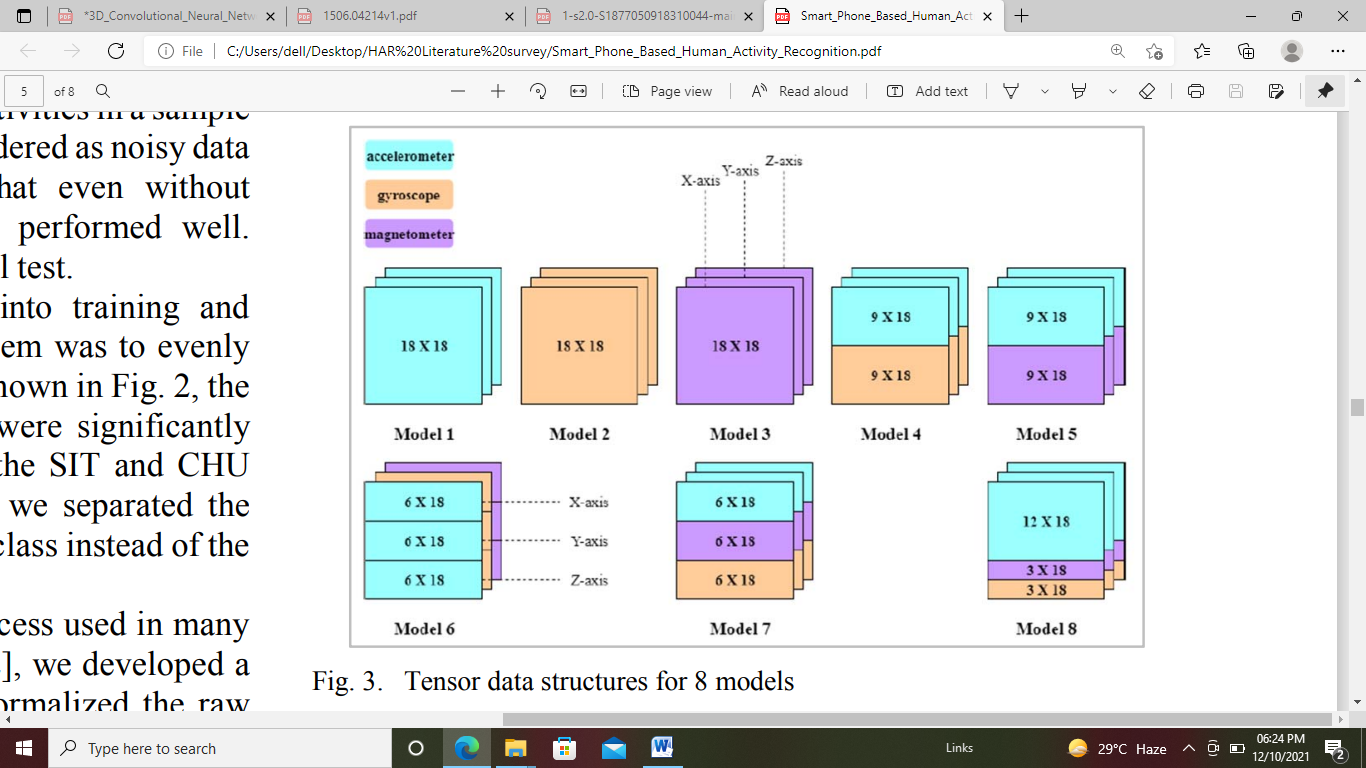


Fig.3. Tensor data structures for 8 models

**E. CNN model**

We created our CNN model based on the architecture which had five sections.

Each of the first two sections contained a convolution layer, a rectified linear unit (ReLU) layer, a max pooling layer, and a normalization layer.

* The convolution layer of a section performed a convolution operation on the outputs of the previous section or the raw inputs. The sequence of layers in each section reduced the total number of weight-variables in the network.
* The ReLU layer exploited an activation function to increase the non-linearity in the feature sets. The max pooling layer extracted the maximum feature map in a set of temporally local neighborhoods.
* The last layer normalized all the features based on the hyper-parameters prior to sending them to the next section.

The third section had a similar structure without the max pooling layer since the dimensions of the feature maps from the convolution layer in the second section were already small enough. The fourth section included a custom fully connected layer followed by the ReLU layer and the normalization layer. The fifth section was a standard fully connected layer.

**Conclusion and future work**

The proposed approach provides a solid foundation for multimodal time-streaming data analysis. Even though the initial goal was to solve the HAR problem, it is possible to extend this approach to any other time series data from the Internet of Things (IoT) to transform raw 2D signals into 3D image patterns. The size of the pre-trained model is less than 10 MB, which is suitable for mobile computing. Nonetheless, no real-time data has been tested with this model, so the performance of this pre-trained model could suffer for noisy data in real-world environment.

**Date: 08/12/21**

**8. An Activity Recognition Framework for Overlapping Activities using Transfer Learning,**

Muhammad Bilal Department of Computer Science, COMSATS University Islamabad, Attock Campus, Pakistan Mubashir Javaid Department of Computer Science, COMSATS University Islamabad, Attock Campus, Pakistan Muazzam Maqsood Department of Computer Science, COMSATS University Islamabad, Attock Campus, Pakistan Seungmin Rho Department of Software, Sejong University, Seoul, Korea Irfan Mehmood Department of Media Design and Technology, Faculty of Engineering & Informatics, University of Bradford, International Conference on Computational Science and Computational Intelligence (CSCI) ,2020.

**Topic:** A Transfer Learning-based Human Activity Recognition (TLHAR) for video data streams.

**Dataset:** UCF101 data set

**Model used:** VGG16 and InceptionV3 (two pre-trained CNN models)

The main contributions to this paper are summarized below:

* Efficiently extract and learn deep features in dynamic video streams containing hidden information.
* Improve activity learning and classification performance on the benchmark dataset.
* Efficiently detect human actions in large video streams and resolve overlapping classes

**Proposed framework**

**A. Dataset**: UCF101 [22] is a publicly available benchmark dataset that has 13320 videos from 101 realistic action classes collected in ‘.avi’ format from YouTube. Each class contains 100∼300 video samples of an action. The duration of each video ranging from 2∼to 7 seconds.

**B. Data Pre-processing:** In data pre-processing, we grouped with those action classes that are similar. After grouping, we extracted frames from each 13320 video and rescaled them to prepare them to input the pretrained model. For VGG16 all the frames are rescaled into

**Date: 09/12/21**

Dimension (224, 224, 3) and for InceptionV3 all the frames are rescaled into dimension (229, 229, 3).

**C. Transfer Learning:** Transfer learning is a technique that allows us to retrain the final layers of a pre-trained model. It drastically reduces dataset requirements and also saves much time for training Deep Neural Network from start. Two of the most famous CNN pre-trained model VGG16 and InceptionV3 are used in this paper. It helps us utilizing these models for our dataset in much less time and with improved accuracy.

**D. VGG-16 CNN model:** A VGG-16 [26] pre-trained CNN model is used in the proposed framework. Pre-trained CNN models learn deep features and are also used for classification. A stack of ‘conv’ layers was used in architecture.

**E. INCEPTION-V3 CNN model**: InceptionV3 pre-trained model is 42 layers deep. The main focus of this model is to use less computational power.

**F. Proposed models:** Taking benefits from Transfer Learning pre-trained VGG16 and InceptionV3 models trained on the ‘ImageNet’ dataset are utilized. The last Fully Connected Layer is removed from the VGG16 pre-trained model. One Flatten Layer is added after removing FCL layers, Dropout of 0.5, ‘relu’ activation, and three dense layers with the SoftMax at the end of the network. InceptionV3 last dense layer is removed and one Flatten Layer is added, Dropout of 0.5, ‘relu’ activation, and three Dense layers with the SoftMax at the end of the network. After some minor pre-processing the frames from the video data are passed to both models. Stochastic Gradient Descent (SGD) is used as an optimizer and for loss; Categorical Cross-entropy is used for both models. At last, both VGG16 and InceptionV3 models are trained with added layers at the end of the model.

**Results and discussion**: A total of 20 classes separated into 6 groups shown in Table I are used in training for both VGG16 and InceptionV3 models. The proposed system was implemented using ‘Python 3’ on Windows 10 environment, Intel(R) Core (TM) i7-8700 set up with 16 GB RAM, and 8 GB NVIDIA GeForce GTX 1070 GPU. For Deep Learning implementation TensorFlow 2.x was used. The proposed system was evaluated using five different evaluation metrics that are: Accuracy, loss, precision, recall, f-measure score.

**Date: 10/12/21**

**9. Real-time Human Activity Recognition Using ResNet and 3D Convolutional Neural Networks**

Archana N., Hareesh K. Dept. of Electronics and Communication Engineering Government College of Engineering Kannur Kannur, 2021.

**Topic:** Real-time human activity recognition by Resnet and 3D CNN without the involvement of the LSTM- attention model.

**Dataset:** Kinetics 400 dataset

**Model:** The combination of Resnet and 3D CNN can enhance the accuracy of recognition.

To improve human action detection accuracy, here the 2D Resnet is converted to a 3D CNN. The kinetics dataset’s large variety of data can help to prevent over fitting during the training phase. Besides, combined Resnet and 3D CNN can improve the recognition accuracy.

**Proposed 3d convolution neural networks and Resnet:**

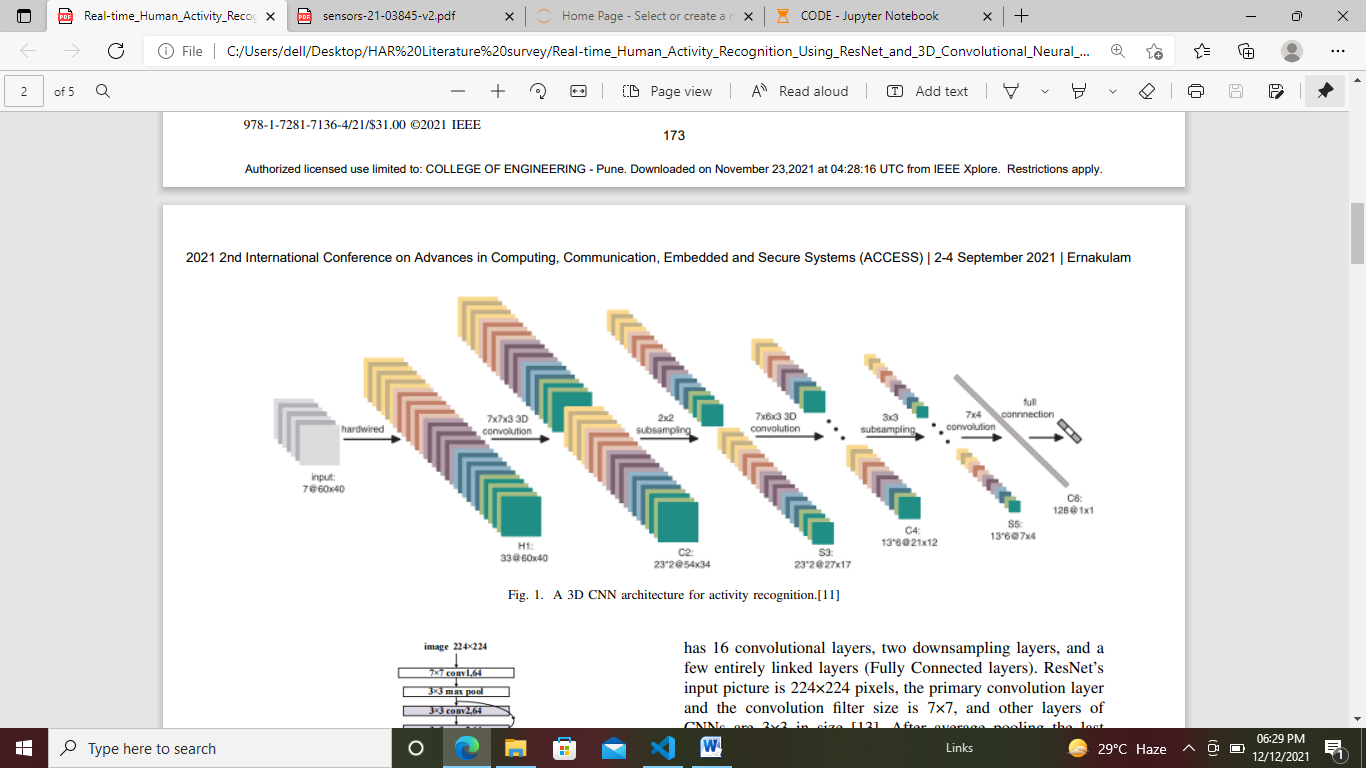
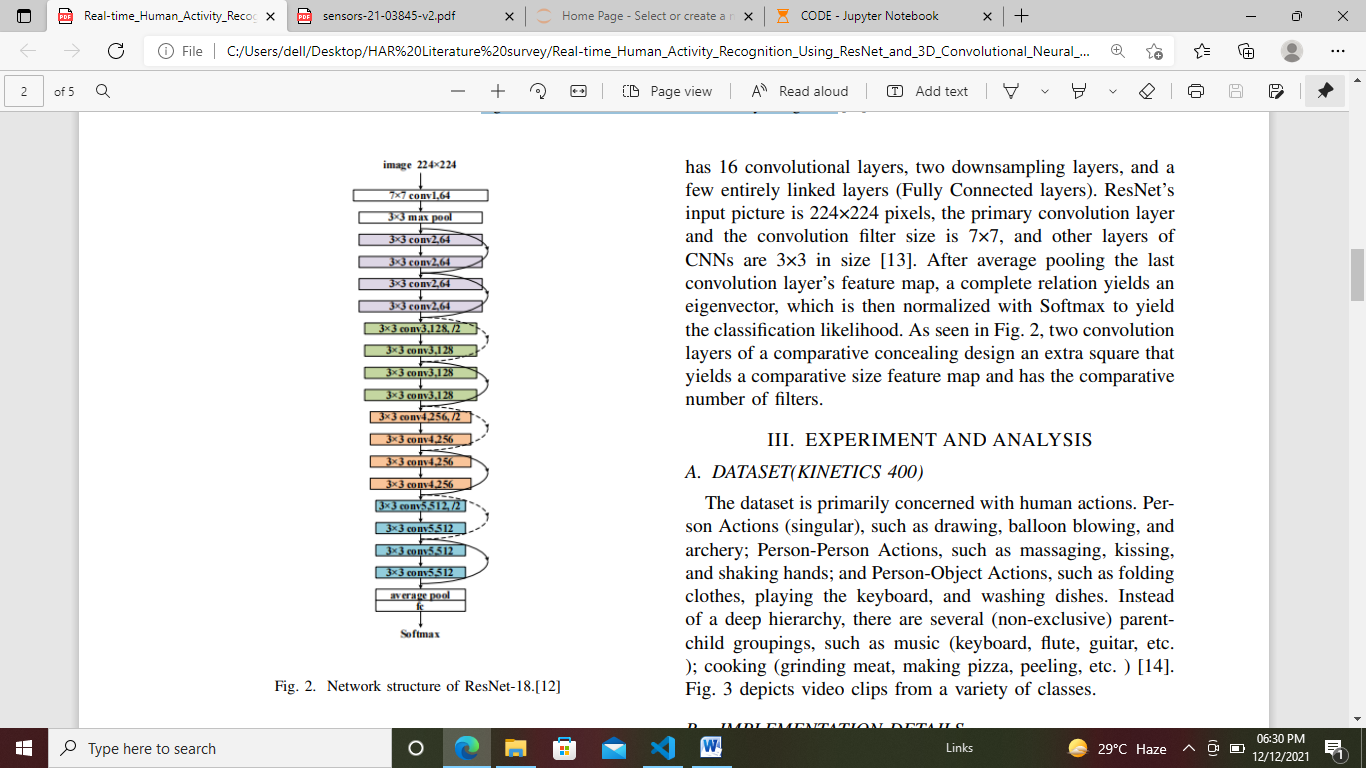
 

Fig.1.A 3D CNN architecture for activity recognition.

Fig.2.Network structure of ResNet-18.

**Experiment and analysis**

**A. Dataset (kinetics 400) :** the dataset is primarily concerned with human actions. Person Actions (singular), such as drawing, balloon blowing, and archery; Person-Person Actions, Dsuch as massaging, kissing, and shaking hands; and Person-Object Actions, such as folding

**Date: 11/12/21**

Clothes, playing the keyboard, and washing dishes. Instead of a deep hierarchy, there are several (non-exclusive) parent child groupings, such as music (keyboard, flute, guitar, etc. ); cooking (grinding meat, making pizza, peeling, etc. ) [14]

**B. Implementation details**: this subsection describes how to train the network for real-time human activity identification on the wide kinetics dataset. The kinetics dataset contains about 400 distinct models of action categories. Then the whole data from the kinetics dataset sectioned into, 80% data for training and 20% data for testing. This is on the grounds that activity location is more intricate if the item utilized for the preparation interaction is tiny. So always use a significant number of samples for the training process than the testing process. To make more data for the training process, data augmentation is used here. It is helpful to make more different data of the same object with a different orientation, scale, etc. It is not an essential step, an optional one. It will help to make the datasets much wider. 3D CNN is capable of extracting spatial and time-related features from raw input. That is, here ResNet-18 network extract both time and appearancedependent information from the relevant frame for action detection. The advanced gradient descent algorithm is used here to train the network. To take a look over to the gradient vanishing problem, the back propagation algorithm is taken and the cross-entropy is considered as a cost function. The network’s rate of learning is fixed as 10−3 . Latter the detection part uses a sliding window system, then the raw video is separated into 16 frames without overlap, then each one is given to the pre-trained network. Then the class score of each frame is found out. The class that has the greatest score demonstrates the distinguished class mark. All tests are implemented based on Tensorflow and Keras.

**Conclusion:** the system that is employed for real-time detection of human motion relay on 3D CNN and ResNet 18. And the developed network is pre-trained on kinetics 400 datasets. The outcome of this work has been concluded that 3D CNN with ResNet-18 can improve the reliability of the network and decrease the losses in training and validation phases. Also, it can be concluded that the use of kinetics 400 datasets avoids the over fitting issue in the network.

**Date: 13/12/21**

**10. An Efficient and Lightweight Deep Learning Model for Human Activity Recognition Using Smartphones**

Ankita 1 , Shalli Rani 1,\* , Himanshi Babbar 1 , Sonya Coleman 2,\*, Aman Singh 3 and Hani Moaiteq Aljahdali, 2021**.**

**Topic:** Convolutional layers are combined with long short-term memory (LSTM), along with the deep learning neural network for human activities recognition (HAR).

**Dataset:**  UCI-HAR for Samsung Galaxy S2

**Model:** Convolutional Neural Network (CNN)- Long short-term memory( LSTM)

Lightweight Deep Learning Classifier Hybrid Model to capture spatial-temporal information in sequential pattern, both CNN and LSTM are being used for recognition with score fusion. Along with this, the experimental results shows the better score fusion between Convolutional network and Long-Short term based memory (LSTM) as compared to LSTM-LSTM.Using this model can not only improve the predictive accuracy but also help in reduction of the models complexity, along with need of advanced feature engineering. This CNN-LSTM can be considered as spatially and temporally deep. This hybrid model effectively covers the problems of image description, video description, and activity recognition. We have proposed the hybrid model for human activity recognition.

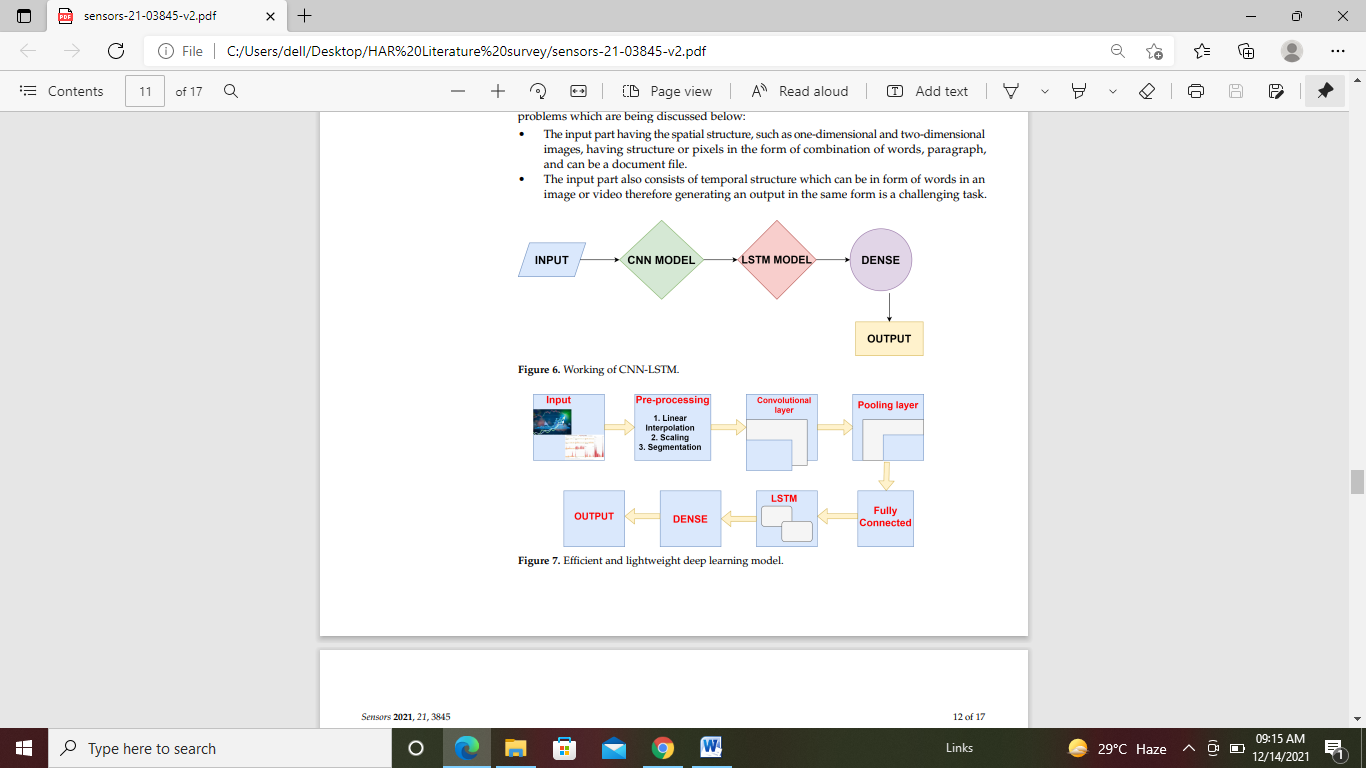
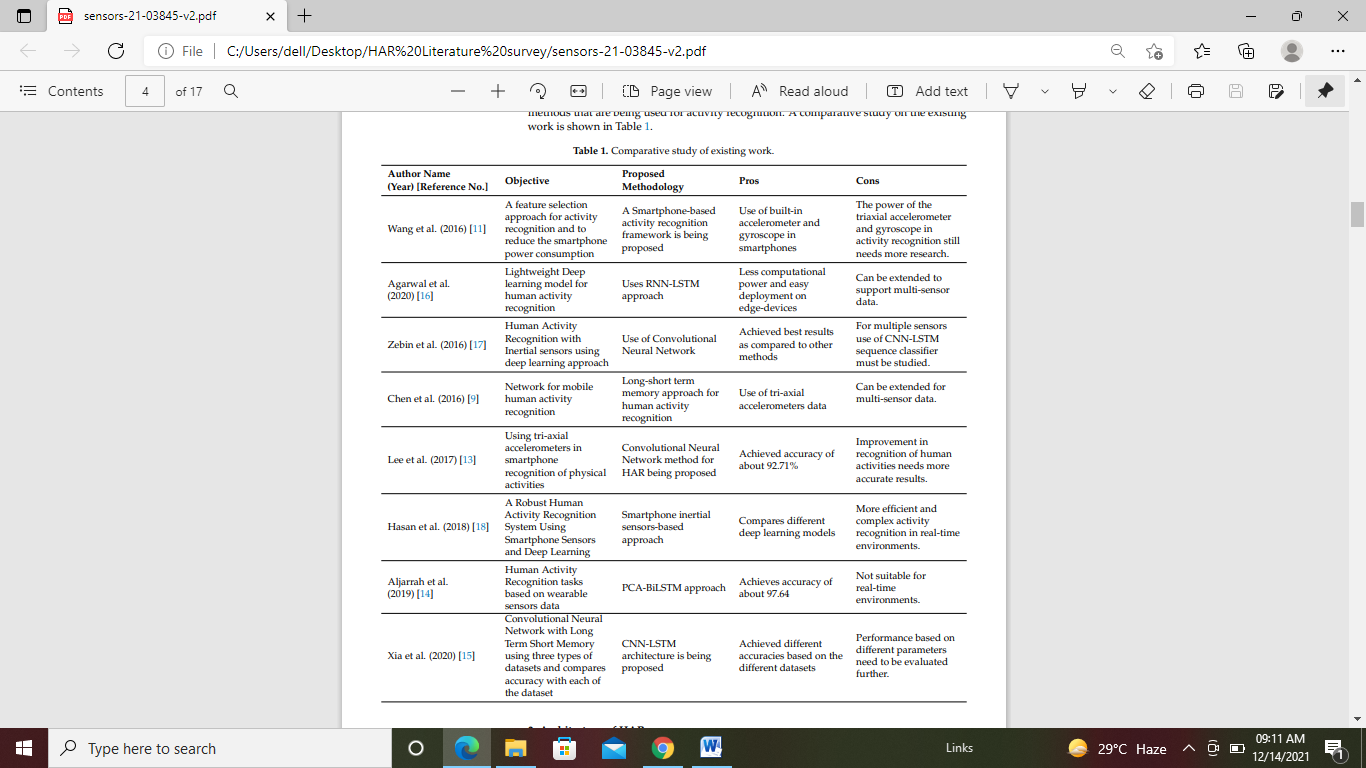


Figure1.Efficient and lightweight deep learning model

Table 1.Comparative study of existing work.



**Architecture of HAR**

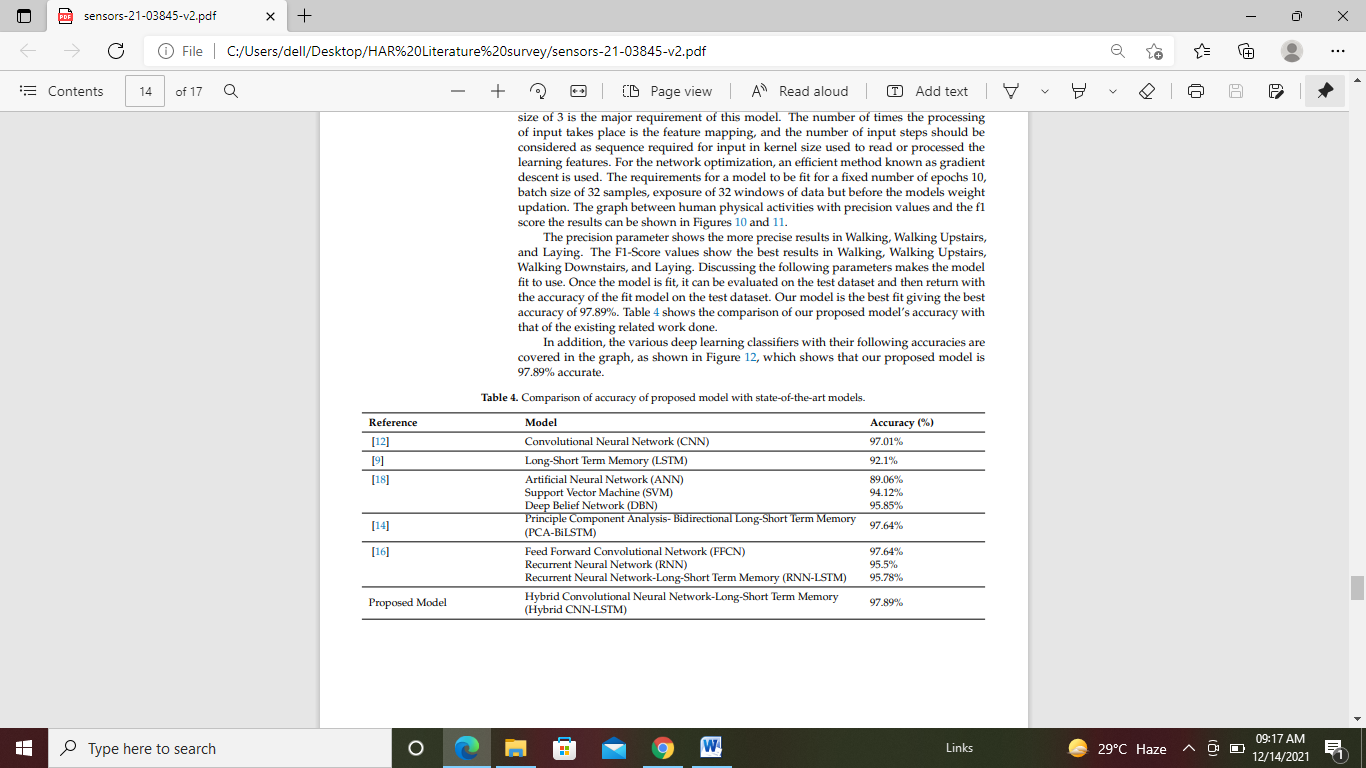
1. Training or Learning Phase:

* Data Collection
* Feature Extraction
* Learning

2. Recognition Phase:

* Data Collection
* Feature Extraction
* Recognition

Table 2.Comparison of accuracy of proposed model with state-of-the-art models.



**Conclusion:**

The proposed model has shown the better accuracy over ANN, LSTM, PCA-BiLSTM, FFNN, DBN, SVM, RNN-LSTM, CNN, and RNN using Gaussian Method. Achieving more accurate results for different types or combinations of deep learning techniques on various types of datasets is the further area of investigation.