ISSN: 0974-5823 Vol. 6 No. 3 December, 2021

International Journal of Mechanical Engineering

VIDEO SEGMENT CLASSIFICATION BASED ON HUMAN ACTIVITY RECOGNITION

Aradhita Menghal, Shubham Jha, Shreyas Nemani, Akshat Chandak, Urvi Negi and Khushboo Khuana*

Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur

Abstract:

Recognizing human activities from video sequences or still images is a challenging task due to problems, such as background clutter, partial occlusion, changes in scale, viewpoint, lighting, and appearance. Many applications, including video surveillance systems, human-computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system. In this work, we provide a detailed review of recent and state-of-the-art research advances in the field of human activity classification. We propose a categorization of human activity methodologies and discuss their advantages and limitations. In particular, we divide human activity classification methods into two broad categories based on data being used from different modalities or not. Then, each of these categories is further analyzed into sub-categories, which reflect how they model human activities and what type of activities they are interested in. Moreover, we provide a comprehensive analysis of the existing, publicly available human activity classification datasets and examine the requirements for an ideal human activity recognition dataset. Finally, we report the characteristics of future research directions and present some open issues present in the human activity recognition task.

Keywords: Human Activity Recognition, Neural Networks

1. INTRODUCTION

Human Activity Recognition (HAR) plays a significant role in human-to-human interaction and interpersonal relations. It provides information about the identity of a person, their personality, and psychological state. It is an intrinsic characteristic of humans to recognize another person's activities. However, it is an intricate task for computers. HAR is the most popular research in the field of computer vision and machine learning. Many applications, including video surveillance systems, human-computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system.

Among various classification techniques two main questions arise: "What action?" (i.e., the recognition problem) and "Where in the video?" (i.e., the localization problem). When attempting to recognize human activities, one must determine the kinetic states of a person, so that the computer can efficiently recognize this activity. Human activities, such as "walking" and "running," arise very naturally in daily life and are relatively easy to recognize. On the other hand, more complex activities, such as "peeling an apple," are more difficult to identify. Complex activities may be decomposed into other simpler activities, which are generally easier to recognize. Usually, the detection of objects in a scene may help to better understand human activities as it may provide useful information about the ongoing event.

The goal of human activity recognition is to examine activities from video sequences or still images. Motivated by this fact, human activity recognition systems aim to correctly classify input data into its underlying activity category. Depending on their complexity, human activities are categorized into: (i) gestures; (ii) atomic actions; (iii) human-to-object or human-to-human interactions; (iv) group actions; (v) behaviors; and (vi) events.

HAR has been a challenging problem yet it needs to be solved. It will mainly be used for eldercare and healthcare as an assistive technology when ensemble with other technologies like Internet of Things(IoT). HAR can be done with the help of sensors, smartphones or images. In this paper, we present various state-of-the-art methods and describe each of them by literature survey. Different datasets are used for each of the methods wherein the data are collected by different means such as sensors, images, accelerometer, gyroscopes, etc. and the placement of these devices at various locations. The results obtained by each technique and the type of dataset are then compared to obtain a result. In this paper, we have presented a system to recognise human activity from input frames of the video.

The rest of the paper is organized as follows. A literature review is presented in section 2. Section 3 walks you through the methodologies and approaches used in this system. It also talks about the part-by-part construction of the model. Later in this section, we discuss the architecture of the model. Section 4 discusses the results we got through several trials and validations. Finally, in section 5 we conclude the paper.

2. LITERATURE REVIEW

To get a proper idea about the human activity recognition system, it was important to review some of the previous works of other researchers in the field. We studied published information in this particular subject area to expand and diversify our knowledge base of this topic.

In (Mutegeki and Han, 2020) a comprehensive deep learning-based activity recognition architecture, a CNN-LSTM is proposed to reduce the complexity of the model and improve the accuracy, in the article (Zeng et al., 2014) a method to draw out discriminative characteristics automatically for recognising activities is presented. The procedure described

here uses CNN and local dependency, as well as scale invariance of signals, is captured, The scene context features that describe the subject's environment at global and local levels are introduced in (Shikha et al., 2020), A DNN structure has been described to acquire human activity's high-level representation merging motion as well as context features. In (Cruciani et al., 2020), using a RESNET-34 3D CNN Model, a HAR system is developed. The model is trained using a Kinetics data set which has 400 classes that depict activities of humans in their everyday life and work and consists of 400 and more videos for each class.

A case study has been presented in (Ann and Theng, 2014) where the use of an already trained CNN feature extractor is evaluated under practical situations. Initially, distinct topologies are evaluated to recognise the best models for human activity recognition. In this way, a pre-trained CNN model is acquired. Then this model is employed in the form of a feature extractor evaluating its use with a large-scale real-world dataset. Recently, thirty-two research papers based on sensing technology that is utilised in human activity recognition is analysed covering majorly three areas: RGB cameras, depth sensors, and wearable devices in (Roshtkhari and Levine, 2013), A detailed description of the pros and cons of the sensing technologies is also discussed.

In paper (Rani et al., 2021), a great method for action recognition, localization, and video matching based on a hierarchical codebook model of local Spatio-temporal video volumes is presented. In (Hammerla et al., 2016) convolutional layers are merged with LSTM, along with the deep learning neural network for human activities recognition. The model described here draws out the features with an automated strategy and then further categorizes them with model attributes. Deep, convolutional, and recurrent approaches across three datasets are discussed in (Sefen et al., 2016) These datasets contain movement data captured with wearable sensors. Off-the-shelf sensors of smartphones and smartwatches are combined and are used for recognizing human activities as proposed in (Weinland et al., 2011) This gives the best tradeoff between the system's computational complexity and recognition accuracy. To achieve this, several evaluations were performed which determined which classification algorithm and features had to be used.

In (Xia et al., 2020) a survey is given which is specifically concentrated on approaches that aim at the classification of full-body motions, such as punching, walking, waving, etc. and they are categorized according to how they represent the spatial and temporal structure of actions. (Hur et al., 2018) proposes a deep neural network that combines convolutional layers with long short-term memory (LSTM) that extracts activity features automatically and classifies them with a few model parameters. An effective HAR method, Inertial sensor signal to Image(Iss2Image) is proposed in (Huang et al., 2019) It is a new encoding technique that transforms an inertial sensor signal into an image with minimum distortion and a CNN model for image-based activity classification. A new method for human activity recognition is introduced in (Khaire et al., 2018) This method has improved recognition accuracy significantly and has reduced complexity. This method uses a two-stage end-to-end CNN and data augmentation.

(Chen and Xue, 2015) presents an approach for activity recognition that is based on ConvNets. This approach combines multiple visual cues. In this paper, skeleton images are created using a new method, from skeleton joint sequences. This represents motion information. In (Lee et al., 2017) a CNN model is constructed to develop an acceleration-based human activity recognition method. Here, the convolution kernel is modified to adapt the characteristics of tri-axial acceleration signals. On the same dataset, a comparison of this method with some methods which are also used to accomplish the recognition is performed. In data of three human activities walking, running, and staying still is gathered from smartphones using a smartphone accelerometer sensor. These human activities are recognized using a 1D CNN-based method.

3. METHODOLOGY:

3.1 DRAWBACKS OF EXISTING WORK

As discussed in the above section many researchers have used the traditional methods to train images based on different actions which lead to inconclusive results and poor accuracy as there are many actions that cannot be firmly decided by the nature of just one frame, consider the action of sitting and standing on a chair both actions seem to be similar only the direction of motion is different i.e the sequence of frame in which action is occurring, Even looking to a single image or frame model cannot predict the action with appreciable accuracy, as there are many factors such as the environment in which the action is taking place is key when it comes to recognizing and many more small factors contribute towards identifying the action to overcome these challenges we propose a single frame CNN approach

3.2 MODELLING APPROACH

The simplest one can think of identifying Human Activities in video segments can be by extracting a single frame of video and then identifying and classifying the activity in that frame. The Human Activity of all the individual frames can be recognized and we can assign the activity to that frame. But there is a drawback of this approach as discussed above, the model will not be able to achieve higher accuracy, another drawback this approach is that if we consider, For example, a person playing football on a field and a person running on a track then the activity of a person playing football can be wrongly classified as running as we are only recognizing key activity. Therefore, there is a need to distinguish between the environmental context and identify.

To tackle this problem, we train the model using multiple images of playing football in field and running on track which would give the model exposure to various environments and then it can give predictions based on considering the environmental factors So, with sufficient examples, the model learns that a person with a running pose on a football field will be playing football and a person with a running pose on track will be running. which would improve the accuracy slightly but not completely as the model will not always be fully confident about each video frame's prediction, so the

predictions will change rapidly and fluctuate. This is because the model is not looking at the entire video sequence but just classifying each frame independently.

The solution to this problem is instead of classifying and displaying results for a single frame, we average results over n finite frames. This would effectively get rid of that flickering. Once we decide the value of n, we can then use the moving average/rolling average technique using an n sized window to achieve this.

Consider,

n - Number of frames to average over

Pf - Final predicted probabilities

P - Current frame's predicted probabilities

P-1 - Last frame's predicted probabilities

P-2 - 2nd last frame's predicted probabilities

.

.

P-n+1 = $(n-1)^{th}$ last frame's predicted probabilities

We will sum the probabilities of all individual recognized activities and take the average of it. The activity with the maximum average is chosen.

For example, If we consider n as 3 and two classes i.e. [Running, Walking] then the predicted probabilities for P-2, P-1 and P are [0.95,0.05], [0.97,0.03], [0.98,0.02] respectively

Then the predicted values, for Running = (0.95 + 0.97 + 0.98)/3 = 0.97, for Walking = (0.05 + 0.03 + 0.02)/3 = 0.03

As 0.97 (Running score) > 0.03 (Walking score), hence, Prediction = Running.

Using this implementation we can get rid of flickering. Apart from averaging the n frames there is also a need to store temporal information of the sequence of frames as the model will not be able to distinguish between a similar set of actions of a person in the same environment as averaging would not be useful here if we do not know the sequence of frames ex: Action of a person sitting down and standing up on a chair will have the same values of average probability. So we also need to consider its sequence.

So, we use Single-Frame CNN Architecture in which we average the probability of n frames and also account for the sequence of frames, utilizing the temporal information in a video to solve the above issues, considering the environmental context. This approach works efficiently as we are averaging n finite frames. Hence we can also take a few spread-out frames out of n frames to avoid unnecessarily classifying all n frames.

3.3. PROPOSED METHODOLOGY

Figure 1 presents the overview of the proposed methodology. Initially, we perform data extraction. To clean and achieve uniformity in the dataset, we preprocess it. Further, we split the data into train and test sets. A sequential model is created with two CNN layers with the activation function as RELU which is discussed further in detail in section IV. We train the model using the training set and then apply the model to the test set to evaluate the performance of our model.

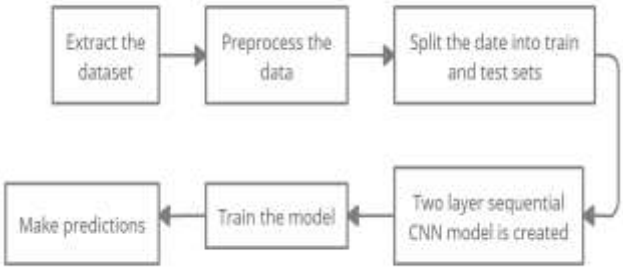


Figure 1. System Overview

Figure 2 describes the architecture of the deep learning model utilized for the task of activity recognition. It consists of 2 consecutive layers of convolutional layers and 9 layers deep with 2 sets of batch normalization which takes inputs as frames of videos and predicts the output.

For the entire video input, a single prediction will take n frames from the entire video and make predictions. Finally, it will average the predictions from those n frames to give us the video's final activity class where the video can comprise of various activities. Now, from this, we extract only that timeline where the user-entered keyword prediction is available.

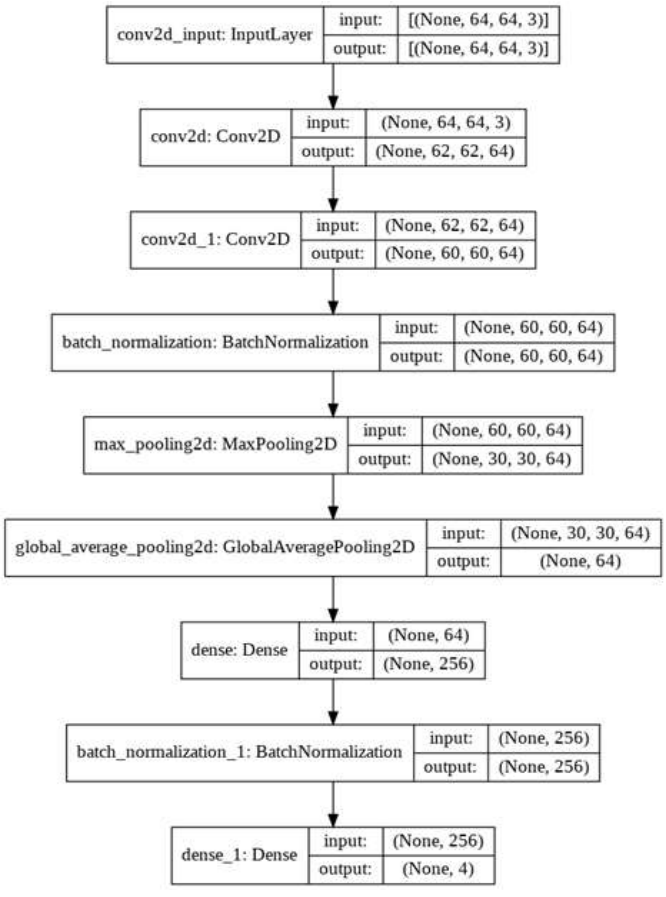


Figure 2. Architecture of Model

3.4 DATASET

The Dataset used in the experimentation is the UCF50 - Action Recognition Dataset. UCF50 is an action recognition dataset that contains 50 Action Categories consisting of realistic videos, 25 Groups of Videos per Action Category,

133 Average Videos per Action Category, 199 Average Number of Frames per Video, 320 Average Frames Width per Video, 240 Average Frames Height per Video and 26 Average Frames Per Seconds per Video.

3.5. DATA PREPROCESSING

A standard set of preprocessing is done to achieve uniformity in the dataset which consists of resizing frames of the video to 64 x 64 and 8000 is set to be the maximum number of training images allowed for each class, After that frames are extracted from each video while performing preprocessing operations like resizing and normalizing images which while reading the video file frame by frame, resizes each frame, normalizes the resized frame, appends the normalized frame into a list, and then finally returns that list.

Next, we perform the following steps using the frame extraction procedure described above and create the final preprocessed dataset.

1. Iterate through all the classes
2. For each class, iterate through all the video files present in it.
3. The frame extraction procedure is applied to each video file.
4. Add the returned frames to a list
5. After all videos of a class are processed, randomly select video frames equal to max images per class and add them to the list
6. Append labels of the selected videos to another list.
7. After all videos of all classes are processed then return both the lists in Numpy array format.

After performing the above steps it returns two lists, which is the final preprocessed data:

- A list of feature vectors
- A list of its associated labels.

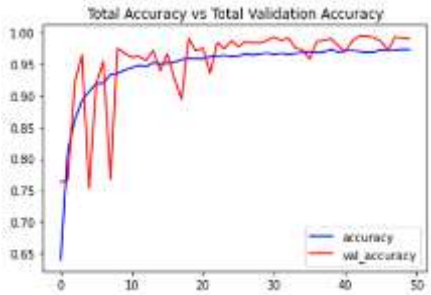
3.6. MODEL CONSTRUCTION

The class labels are converted to a one-hot encoded vector, having two NumPy arrays, one containing all images. The second one contains all class labels in one hot encoded format. The data is divided to create a training and a testing set, a sequential model is created with two CNN layers with the activation function as RELU, below flowchart shows the model's structure and layers

4. RESULTS AND DISCUSSION:

We have used Single-Frame CNN Architecture with a sequential model having two CNN layers with the activation function as RELU. To predict the output we have taken an average of n frames. The model is giving an accuracy of 99%. We are getting the best accuracy around 50 epochs. The accuracy is increased by increasing the number of epochs. Initial accuracy was around 65%. Also, the initial loss was more than 80% but as we increased the epochs the loss decreased to around 10%. Figure 3 shows the loss and accuracy graphs. Further, figure 4 shows the input video along with the frame activity result and video class label.

(a) Total Loss vs Total Validation Loss



(b) Total Accuracy vs Total Validation Accuracy

Figure 3. Graph based evaluation

(a) Input Example

(b) Output Example

Figure 4. Sample Results

5. CONCLUSION:

Human Activity Recognition (HAR) provides information about the identity of a person, their personality, and psychological state. It plays a significant role in human-to-human interaction and interpersonal relations. In this paper,

we looked into the drawbacks of existing methods of human recognition. To overcome these drawbacks, instead of using a single frame, we proposed to use multiple frames and take their averages to find the activity label. This procedure is efficient because we are averaging n frames and considering temporal storage. Hence, we have successfully presented a structured approach for the HAR system. In the future, we would like to integrate our proposed method with applications used for smart surveillance systems, medical research, automatic sports commentary, and others.

6. REFERENCES:

Ann, O.C. and Theng, L.B., 2014, November. Human activity recognition: a review. In 2014 IEEE international conference on control system, computing and engineering (ICCSCE 2014) (pp. 389-393). IEEE.

Chen, Y. and Xue, Y., 2015, October. A deep learning approach to human activity recognition based on single accelerometer. In 2015 IEEE international conference on systems, man, and cybernetics (pp. 1488-1492). IEEE.

Cruciani, F., Vafeiadis, A., Nugent, C., Cleland, I., McCullagh, P., Votis, K., Giakoumis, D., Tzovaras, D., Chen, L. and Hamzaoui, R., 2020. Feature learning for human activity recognition using convolutional neural networks. CCF Transactions on Pervasive Computing and Interaction, 2(1), pp.18-32.

Hammerla, N.Y., Halloran, S. and Plötz, T., 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. arXiv preprint arXiv:1604.08880.

Huang, J., Lin, S., Wang, N., Dai, G., Xie, Y. and Zhou, J., 2019. TSE-CNN: A two-stage end-to-end CNN for human activity recognition. IEEE journal of biomedical and health informatics, 24(1), pp.292-299.

Hur, T., Bang, J., Lee, J., Kim, J.I. and Lee, S., 2018. Iss2Image: A novel signal-encoding technique for CNN-based human activity recognition. Sensors, 18(11), p.3910.

Khaire, P., Kumar, P. and Imran, J., 2018. Combining CNN streams of RGB-D and skeletal data for human activity recognition. Pattern Recognition Letters, 115, pp.107-116.

Lee, S.M., Yoon, S.M. and Cho, H., 2017, February. Human activity recognition from accelerometer data using Convolutional Neural Network. In 2017 ieee international conference on big data and smart computing (bigcomp) (pp. 131-134). IEEE.

Mutegeki, R. and Han, D.S., 2020, February. A CNN-LSTM approach to human activity recognition. In 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC) (pp. 362-366). IEEE.

Rani, S., Babbar, H., Coleman, S., Singh, A. and Aljahdali, H.M., 2021. An Efficient and Lightweight Deep Learning Model for Human Activity Recognition Using Smartphones. Sensors, 21(11), p.3845.

Roshtkhari, M.J. and Levine, M.D., 2013. Human activity recognition in videos using a single example. Image and Vision Computing, 31(11), pp.864-876.

Sefen, B., Baumbach, S., Dengel, A. and Abdennadher, S., 2016, February. Human activity recognition. In Proceedings of the 8th International Conference on Agents and Artificial Intelligence, SCITEPRESS-Science and Technology Publications, Lda (pp. 488-493).

Shikha, M., Kumar, R., Aggarwal, S. and Jain, S., 2020. Human activity recognition. International Journal of Innovative Technology and Exploring Engineering, 9(7), pp.903-905.

Weinland, D., Ronfard, R. and Boyer, E., 2011. A survey of vision-based methods for action representation, segmentation and recognition. Computer vision and image understanding, 115(2), pp.224-241.

Xia, K., Huang, J. and Wang, H., 2020. LSTM-CNN architecture for human activity recognition. IEEE Access, 8, pp.56855-56866.

Zeng, M., Nguyen, L.T., Yu, B., Mengshoel, O.J., Zhu, J., Wu, P. and Zhang, J., 2014, November. Convolutional neural networks for human activity recognition using mobile sensors. In 6th international conference on mobile computing, applications and services (pp. 197-205). IEEE.