iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VIII **Month of publication:** August 2020

DOI: https://doi.org/10.22214/ijraset.2020.31009

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VIII Aug 2020- Available at www.ijraset.com

Human Action Recognition in Videos

Kavya Tolety[1], T Srinivasa Rao[2]
[1]*B.E. Student,* [2]*Professor, Department of Computer Science and Engineering, Sambhram Institute of Technology*

Abstract: ***The goal of this project is to develop a Human Action Recognition system with a video classification approach. Data is gathered by downloading images from google, analyze the frames of the video and predict the actions being performed. This is achieved by using Computer vision - a field of computer science that works on enabling computers to see, identify and process images similar to human vision, and then provide appropriate results. It is like imparting human intelligence and instincts to a computer.***

Keywords: Human Action Recognition, Moving average prediction, transfer learning, video classification, fine tuning

I. INTRODUCTION

Action recognition can be defined as the identification of a set of predefined action classes when a video is designated. It is required to deduce the Spatio-temporal location where the action occurs. The main problem in action recognition is to locate the exact locus as to where the movement occurs. In addition to this, we have many other complications like occlusion, spatial complexity, and background clutter.

Computer vision is a field in scientific research that deals with how computers analyze and gain understanding from visual images or videos. It understands and automates functions that can be attained by the human visual system. The trained system analyzes the digital world. Systems can use images from webcams, videos, and models to identify, classify, and eventually perform operations based on the inspection. Information can be extracted from image data to figure out its properties using various models. Based on Learning theory, statistics, physics, and math models are built. The information extracted can be of several forms, for instance, views from cameras, or video sequences.

Convolutional Neural Networks is a powerful deep learning technique which preserves the spatial structure of the problem. They are popular because state-of-the-art results can be achieved on difficult computer vision and natural language processing tasks.

Video content analysis is a domain of deep learning with several applications, one of which is human action recognition. Its goal is to recognize activities from a sequence of surveillance on the actions of matter and the surrounding environment. Human action recognition is a complex technology due to the difficulty to extract information about a person's identity and their psychological states. Hence, human action recognition has gained the interest of researchers in the computer vision area and many approaches have been proposed for the aim of developing this technology.

Human action recognition approaches can be divided into these categories as shown in Figure 1:

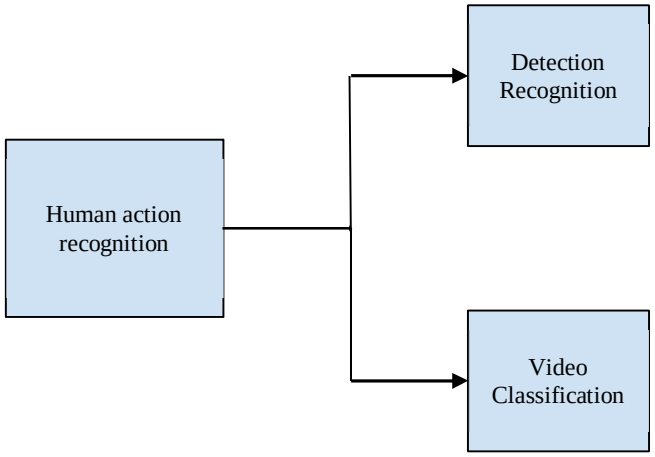


Figure 1: Human Action recognition classification

1) *Detection Recognition:* Start by first detecting the motion of the person's action, followed by recognizing the action(s).
2) *Video Classification:* Videos are classified according to the action a video contains.

In this paper, we are mainly going to focus on the video classification approach for HAR.

There are several applications of action recognition systems. Action recognition plays a major role in many fields and applications, some of them are:

a) *Video Monitoring:* Video surveillance or monitoring is the task of observing humans and objects of particular view using cameras. Surveillance systems are required for safety reasons they help to detect suspicious activities in banks, ATMs, or other public places to avoid burglary and malicious activity. The increase in demand for these installed systems are mostly in big cities or towns, hence there is a need for action recognition systems as well.

b) *Medical and Health care Industry:* The demand for systems that recognize actions and detect early symptoms of physical or mental problems of patients or older people is increasing rapidly. The task is to identify and understand behavioral changes like posture, sleeping habits, and so on. It is a necessity for older people who live alone without much safekeeping or supervision.

c) *Human-Computer Interaction and Entertainment Industry:* It is a very generic thing for humans to interact directly with a video device for communication or interaction purposes. In the entertainment industry, these are used by a large number of people for gaming, sensors, etc. The most popular example is the Microsoft Kinect.

II. RELATED WORK

A survey [1] which talks about the various techniques developed and the different angles of approaches in the field of research which aims to recognize human actions from videos. It broadly classified Human action recognition approaches into two categories: Action detection method and video classification approach. Unlike other surveys that discuss approaches for action modeling and classification, it summarizes methods for recognizing interaction between humans and object and action detection methods.

Karen Simonyan [2] proposes a method which utilizes a two-stream Convolutional neural network architecture, consisting of two components: (a) Spatial network - works on recognizing actions from still images. (b) Temporal network - the input is formed by stacking flow representations between numerous consecutive frames. Input configurations such as Optical flow stacking, Trajectory stacking, Bidirectional optical flow, mean flow subtraction are used. To see an increase in the quantity of the training data and to better the performance, multitask learning can be used. Separate recognition streams have been incorporated based on ConvNets. Presently, it is seen that temporal networks perform better than their spatial counterparts.

AJ Piergiovanni [3] shows how a convolutional layer that inclines towards optical flow algorithms is used to gain knowledge of motion representations. The flow layer is designed in a way through which it captures the stream of a representation channel inside a CNN for recognizing action. A representation flow layer which can be trained is introduced, using optical flow algorithms. After differentiating between various forms of the layer, it is confirmed that the repetitive optimization and trainable parameters are imperative. The model is more efficient than existing models in terms of speed and accuracy. The notion of 'flow of flow' to evaluate consistent motion representations and indicated that its benefits performance.

Chi Geng [4] shows how CNN is used for action recognition, which acts directly on the inputs. It is made up of two components- a CNN, which behaves as a feature extractor and SVM, which helps in recognizing patterns. In order to enable better real-world applications and lower the already high computational cost of training, a more robust pre-training technique has been introduced. The chosen approach evaluated on the KTH dataset achieves better results as compared to more developed algorithms that use hand-designed features.

Here, Chi Geng [5] shows how a pre-trained CNN model can be used as an approach for detecting human actions. The above-mentioned model is used for feature extraction & representation. For classifying actions, a combination of SVMs and KNNs is used. The features and representations learned by the pre-trained model trained on annotated datasets like ImageNet can be transferred to new tasks with action recognition. The evaluation of the proposed method is performed on benchmark action datasets - UCF sports and KTH. This evaluation shows that the proposed method achieves greater accuracy over other methods.

III. METHOD

With a rise in demand for Human Action Recognition, we aim to create a system that can identify actions being performed by an individual dynamically. Computer vision is used to achieve this prediction and hence estimate accuracy of the actions. Computer vision is a field of research that works on enabling computer systems to observe, analyze and process images in the similar way fashion that human vision does and then provide required output.

The following are the steps for the current action recognition frameworks: a) Feature extraction; b) Dictionary learning, to depict a video based on the extracted features; c) Representation; to classify actions.

In the proposed system, different actions are identified from video clips (which is a sequence of images). The action may not be performed in the entire duration of the video. This is a natural extension of image classification from a single frame to multiple frames and eventually aggregating the prediction from each frame. To smoothen the predictions, a moving average prediction algorithm is used.

In Figure 2, we represent the architecture which identifies different components and represents it graphically providing an overview of the entire system.

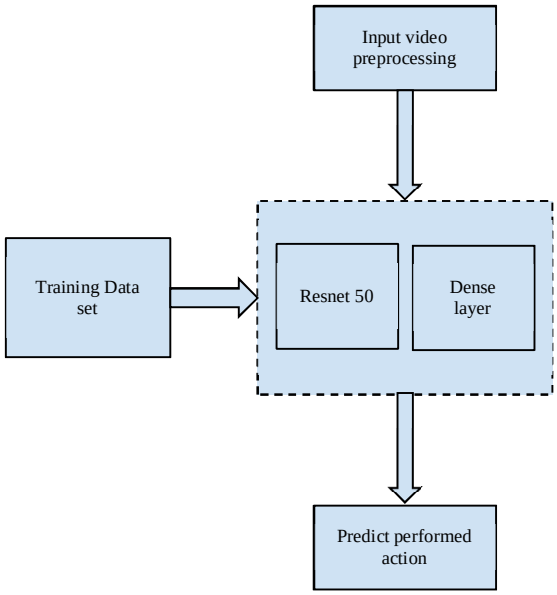


Figure 2: System architecture

The proposed architecture consists of two essential modules:

1) *Resnet50 Model:* is a convolutional neural network that is 50 layers deep. From the ImageNet database one can load a pre-trained version of the network which can classify images into a thousand different classes and since it was trained using more than hundred thousand images.
2) *Fully-Connected/dense layer with SoftMax Activation:* dense layer is a regular deeply connected network in which each input neuron is connected with the output neuron. SoftMax layer converts the output of the network's penultimate layer into a probability distribution and provides the confidence score for each input class.

IV. EMPIRICAL EVALUATION

Dataset There is a huge variety of benchmarks for human action recognition. We make our own dataset from Google by writing a code to grab all the URLs of the images we want to download into a text file. Once the text file containing all the URLs is generated, use code to download the images. Dataset consists of 415 images in each of the 5 different action classes: 1) Clapping 2) Sitting 3) Walking 4) Waving

Training Images classifier is built by loading the dataset and each image in the dataset is pre-processed by resizing it to 224x224 and then performing ImageNet mean subtraction (subtraction in RGB order). Here we use Fine-tuned ResNet50 to classify the images. Fine tuning involves various steps: 1) Initially, Remove FC node at the end of the pre-trained Resnet50 network 2) Replace it with freshly initialized ones 3) Freeze the earlier CONV layers in the network 4) Start training by only training the FC layer heads. Some of the parameter specifications are: 1) Learning rate: 1e-4 2) Epochs: 50 3) Decay: Learning rate/no. of epochs 4) Momentum: 0.9 5) Dense units: 512 6) Number of classes: 8 7) Batch size: 32

Prediction and visualization Using an algorithm called rolling averaging over predictions, the image classifier built in the training step is converted into a video classifier as given in the figure 3. First the Frames from the input video are extracted and pre-processed. The trained model is then fed with each frame extracted from the video. The predictions on each frame obtained from the classifier is appended to a deque. The average of a predetermined number of predictions is computed and the label with the largest corresponding probability is chosen. Finally, Label the frame and write the output to disk.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VIII Aug 2020- Available at www.ijraset.com

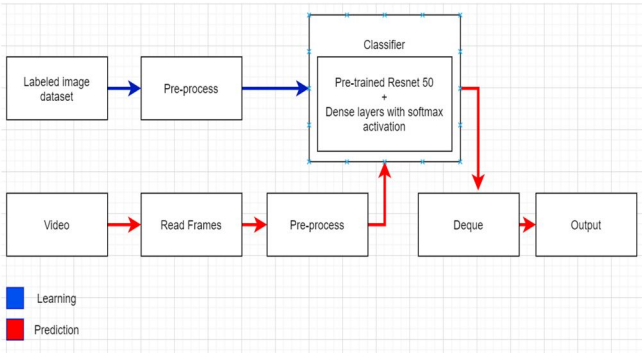


Figure 4: Dataflow diagram

V. RESULTS

Confusion matrix also known as error matrix, is a table layout used to visualize/describe the performance of classification models for a set of test data for which the true values.

The classification report is a representation used to measure the quality of predictions from the classification algorithm. The classification report for the algorithm used in this project is shown below in figure 4.

	precision	recall	f1-score	support
clapping	0.78	0.67	0.72	104
sitting	1.00	0.58	0.73	104
standing	0.45	0.67	0.54	103
walking	0.61	0.58	0.59	104
waving	0.50	0.56	0.53	102
accuracy			0.61	517
macro avg	0.67	0.61	0.62	517
weighted avg	0.67	0.61	0.62	517

Figure 4: Classification Report

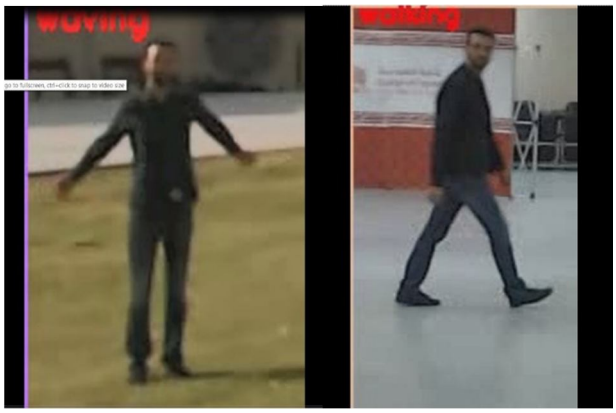


Figure 5: Snapshots of the output

VI. CONCLUSION

A video classification approach in which the frames of a video are considered to be independent of each other can cause something called Label Flickering. This flickering occurs when the trained model gives back different labels for each frame, even if the frames are expected to have the same labels.

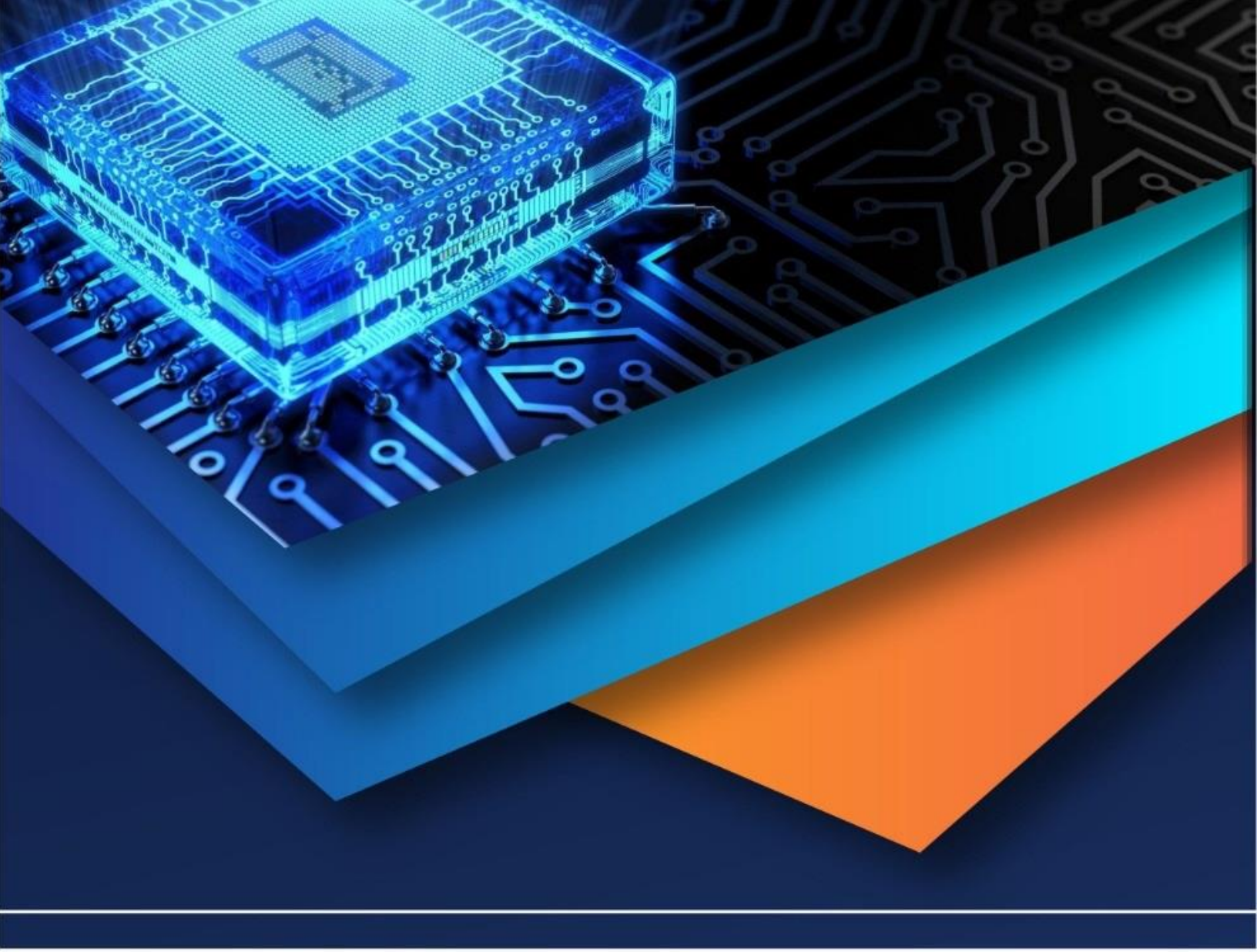
In this paper we have proposed a simple but effective solution to this problem - moving prediction averaging - has been identified and implemented using Keras, TensorFlow, deep learning, and Computer vision. This method involves feeding each frame of the video into a trained model. Then, the predictions on each frame obtained from the classifier are stored in a double-ended queue.

Finally, the average of a predetermined number is computed, and the frames are labelled with the largest corresponding probability. Here, the temporal nature of videos is taken into consideration by making an assumption that frames of a video occurring or coming later in time have some sort of a mutual connection with respect to their semantic objects.

Therefore, the moving prediction averaging algorithm helps in smoothing out the predictions eliminating Label Flickering in the video classification approach. As future work, we aim to carry out the following: 1) Collect more data in class to increase accuracy. 2) Train the model to identify other action classes like human-object interactions and human-human interactions.

REFERENCES

[1] A Comprehensive Survey of Vision-Based Human Action Recognition Methods.2019, Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du and Duan-Sheng Chen
[2] Two-Stream Convolutional Networks for Action Recognition in Videos.2014, Karen Simonyan and Andrew Zisserman
[3] Representation Flow for Action Recognition.2018, AJ Piergiovanni and Michael S. Ryoo, Lehigh University
[4] Human Action Recognition based on Convolutional Neural Networks with a Convolutional Auto-Encoder.2016, Chi Geng, JianXin Song
[5] Human Action Recognition using Transfer Learning with Deep Representations, 2017, Chi Geng, JianXin Song.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY