Turkish Journal of Computer and Mathematics Education Vol.12 No.11 (2021), 818-824

Research Article

Action Recognition Using Deep Learning And Cnn

Mihir Verma[1], Palash Sharma[2], Dr. M. Baskar[3]

[1]Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Chennai, Tamilnadu, India-603 203.

[2]Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Chennai, Tamilnadu, India-603 203.

[3]Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Chennai, Tamilnadu, India-603 203.

[1]md3958@srmist.edu.in ,[2]pa2538@srmist.edu.in ,[3]baashkarcse@gmail.com*

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: Automated action recognition using Deep learning and CNN is playing a vital role in today's day to day society, it may be video action recognitions through cctv, or it may be the smart homes. Now day's human actions are used in many devices to control them like HoloLens VR, for that recognition of action is important that why video recognition. This Paper represents practical, reliable, and generic systems for video-based human action recognition, technology of CNN network is used to recognize different layers of the video images features. These features are obtained by extracting the features from different layers that are through the CNN (Convolutional Neural Network).

Keywords: CNN, Image Processing, Deep Learning, Visual Recognition, Prediction, Artificial Intelligence, Machine Learning.

1. Introduction

Human activity identification is becoming increasingly relevant, not just in terms of security and surveillance, but also because of psychological interests in understanding human behavioral patterns. This report is a review of numerous current techniques that have been brought together to form a working pipeline for the study of human interaction in social gatherings. Action Recognition and Security System Using any security camera devices or recorded surveillance and Integration using ML Algorithm is to ease the process of detecting patterns and to make a prior decision on what action is being performed in the surveillance.

We additionally thought on recognising action like fighting, running and jogging to get extra details through the surveillance data. This model could be useful for modern purpose on assessing the security footage and increasing home security and various high security environments. The term digital image refers to process a 2-dimensional image by a digital pc. It implies digital processing of any two dimensional data. A digital image is an array of real or complicated varieties diagrammatic by a finite number o fbits.

2. Related Works

The researchers have discussed several techniques in action recognition. This section explores different approaches towards the problem.

The author has stated in [1][10] LTRCN (Long-term Recurrent Convolutional Networks) for Visual Recognition, we describe a class of recurrent Convolutional neural layer architecture which is end-to-end trainable, suitable for big-scale visual understanding tasks, and shows the value of these models for activity recognition, image captioning, and video description.

In the following [2][9] the author discussed about Observing Human and Object's Interaction between them: It Uses Spatial and Functional Compatibility of the device for Recognition, it includes getting scene/occasion, breaking down human developments, perceiving manipulable articles, and noticing the impact of the human development on those items. While every one of these perceptual errands can be directed autonomously, the acknowledgment rate improves when cooperation's between them are thought of.

In this study [3][11] the author has discussed Visual Event Recognition in Videos by reading from web data, we propose a visual event for the visual space of consumer space recording using many indirectly marked web recordings (e.g., from YouTube). First, we propose another time-adjusted pyramid guided by the process of measuring the distances between two video cuts, in which all video cut is limited to time-varying volumes at different levels.

The author has stated in [4] about Job Recognition Using Non-Opposition Representation and General Basic Selection, There are three stages connected to our approach. First, we suggest another adjective to remember location, called private word setting, to improve the stigma of nearby descriptions that use location. Second, from an understanding of the speculative set definition, we study the work units using a non-contradictory visual chart that makes a section presentation and incorporates mathematical details.

In the following [5] the author talks about Slow Fast Networks for Video Recognition, We present Slow Fast networks for video acknowledgment. Our model includes (I) a slow pathway, working at low edge rate, to catch spatial semantics, and (ii) a Fast pathway, working at high edge rate, to catch movement at fine fleeting goal. The Fast pathway can be made lightweight by decreasing its channel limit, yet can learn valuable fleeting data for video acknowledgment.

In [6][12] the author discussed Late Temporal Modeling in 3D CNN Architectures with BERT for Action Recognition, we join 3D convolution with late worldly displaying for activity acknowledgment. For this point, we supplant the ordinary Temporal Global Average Pooling (TGAP) layer toward the finish of 3D Convolutional engineering with the Bidirectional Encoder Representations from Transformers (BERT) layer to more readily use the fleeting data with BERT's consideration component.

The author stated in [7] about Towards Fast Action Recognition by means of Learning Persistence of Appearance, efficiently displaying dynamic movement data in recordings is essential for activity acknowledgment task. Best in class strategies vigorously depend on thick optical stream as movement portrayal. Despite the fact that joining optical stream with RGB outlines as info can accomplish great acknowledgment execution, the optical stream extraction is very tedious.

In this [8][13]the author talks about Mutual Modality Learning for Video Action Classification, The development of models for video activity characterization advances quickly. Be that as it may, the presentation of those models can in any case be effectively improved by assembling with similar models prepared on various modalities (for example Optical stream). Shockingly, it is computationally costly to utilize a few modalities during deduction.

3. Proposed system

Convolutional Neural Networks (CNN) is one of the first popular algorithmic projects for learning and is also widely used in image editing systems. When all is said and done, CNN configuration consists of 3 layers, which are Convolutional layers, cohesive layers, fully connected layers. CNN counts detect image detail across layers to see alternatives and image detection, so it provides a compilation effect. The CNN design consists of incorporating layers of Convolutional and cohesive layers, followed by a collection of fully compatible layers. The product of each layer within CNN is the next layer offering. CNN contribution can be 3D image (width x length × width), scaling and moreover, stem size images. The most common is the integration of information channels with its 3 red, green, and blue (RGB) channels. Convolutional layers separate selections from images. Each Convolutional has multiple networks known as channels or sections that run over the image information to get the visual information in the image. CNN's key layer channels distinguish simple colors and patterns. At that point within the next layers, they gradually separated the larger designs. Outstanding search, each channel uses the convolution function to generate an object map.

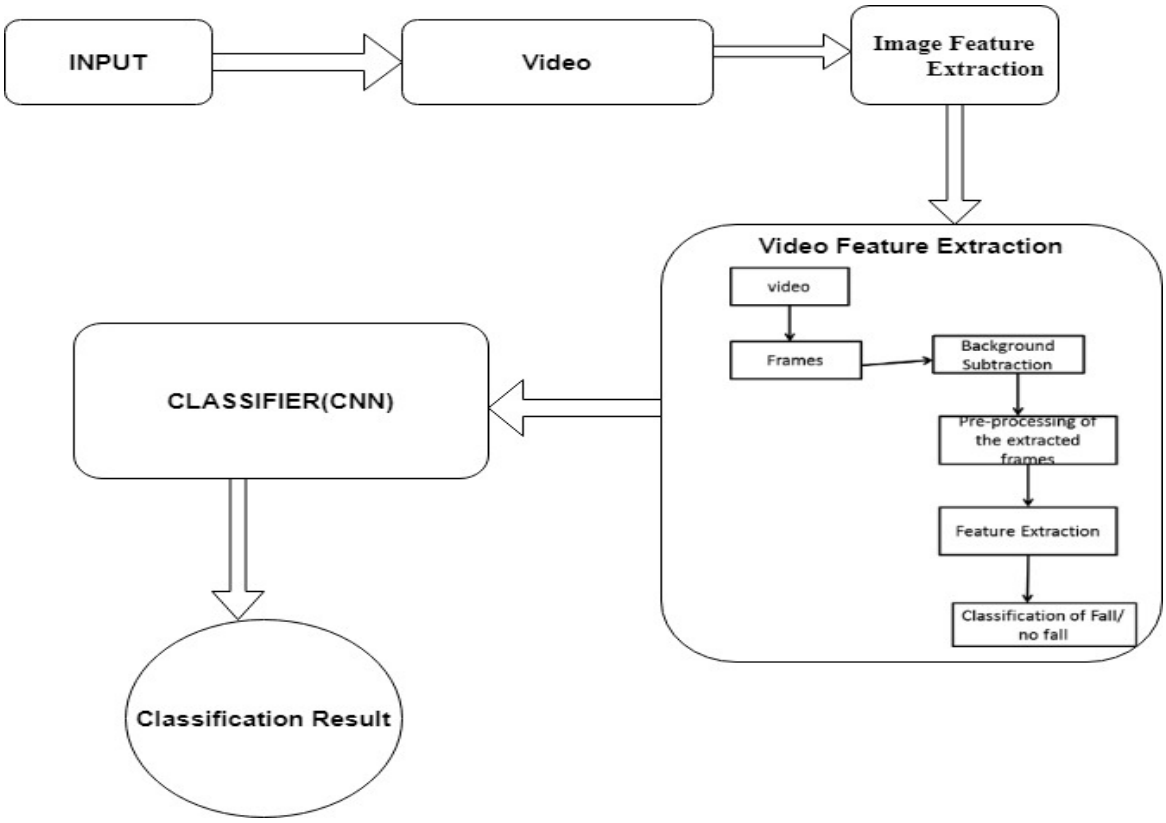

Figure 1. Architecture of action recognition using CNN

The functional architecture of proposedaction recognition using CNN based approach is presented in Figure 1

4. Train the dataset

Generation of the training set and test data set is as far as a basic data division. You should initially isolate the n-classes of data. This data would then be able to be split into a basic 70-30 or down the middle proportion of training, test, and validation data set. Details can be arranged so that all classes have equivalent portrayal in all evaluation and training The Feature vector is only an assortment of data of qualities or highlights in a picture. In the event of Photos, they might be mathematical highlights, and surface highlights, and so forth presently, for AI you should give the training data to the classifier with the goal that it can fabricate a model. This model can be checked and the proper portion boundaries are chosen. K-overlap cross approval is perhaps the least complex approach to pick something similar. There are likewise a few calculations you can investigate for choosing great (mRmR) highlights. When your highlights and model have been made you can test the detachment and data and check the exactness of the partition.

5. Classification

Convolutional Neural Networks (CNN) is one of the most in-depth and widely used learning algorithms in image classification systems. Generally, CNN formats have three types of layers, namely Convolutional layers, composite layers, and fully integrated layers. The CNN algorithm detects the input image that it transmits to the layers the separation involves visualizing the image, and then creates the effect of the collection into various vectors of harvest. The CNN design consists of flexible layers and reconciliation issues, followed by a set of fully integrated layers. The production of each layer on CNN is a matching layer offering. CNN contribution to 3D image (width x height x depth), width and height are the elements of the included images. Multiple multi-channel channels with three red, green, and blue (RGB) channels. Convolutional layers produce highlights in the images. Each modification has metric loads that are considered to be channels or characters that go above the image information to get a clearer picture. CNN's main layers distinguish tones and specific examples. At the same time, in the accompanying layers, they gradually found more complex patterns. An outstanding finding, each channel uses the help of convolution to split an object map.

6. Results and discussions

In this research the experimentation were conducted using the MATLAB programming language. We further used the Image Processing Toolbox for cropping the image component we were taking from the video and for the analysis of that image we used the Deep Learning Toolbox. Both of these are the widely used tools of the MATLAB and due to their precise results we took both of these in our experiment. In the table 1 we have discussed some of the algorithms and their accuracy.

Model	Accuracy
CNN only(BVLC-caffenet)	19.32
CNN only(Resnet-50)	63.72
super-class-model	67.30
next-frame-prediction	68.20

Table 1. Algorithms and their accuracy

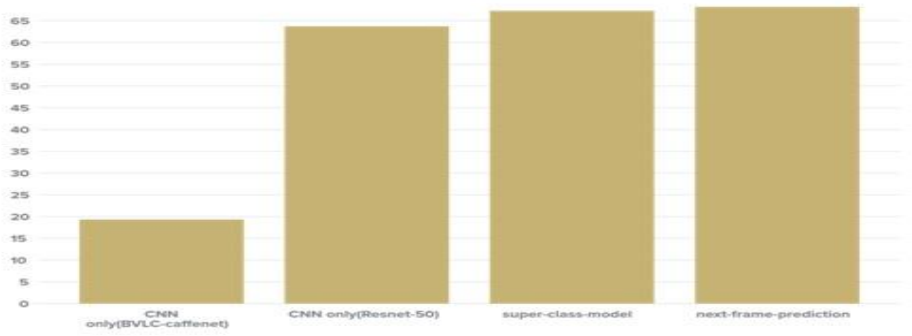

Figure 2. Analysis of algorithms and their accuracy

Some other STATE-OF-THE-ART algorithms and their results and accuracies discussed in the table 2:

Model	Accuracy(%) on UCF-101
Slow Fusion CNN [9]	65.3%
LRCN [1]	82.3%
ActionFlowNet [12]	84.1%
Composite LSTM [16]	84.2%
Two stream ConvNet [13]	88.1%
LSTM + CNN (Optical Flow + Image Frames) [18]	88.5%

Convolutional Two stream fusion [3]	92.4%
Convolutional Two stream fusion+IDT [3]	93.3%
ST-ResNet [2]	93.3%
ST-ResNet+IDT [2]	94.5%
(Ours)next-frame-prediction	68.2%

Table 2-Some other state of the art algorithms and their accuracy

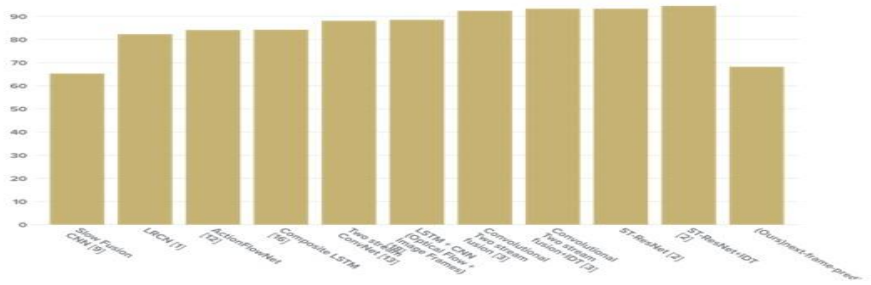

Figure 3. Analysis of some other state of the art algorithms and their accuracy

Based on our research and after executing the code in the MATLAB we got the following results which is shown below images

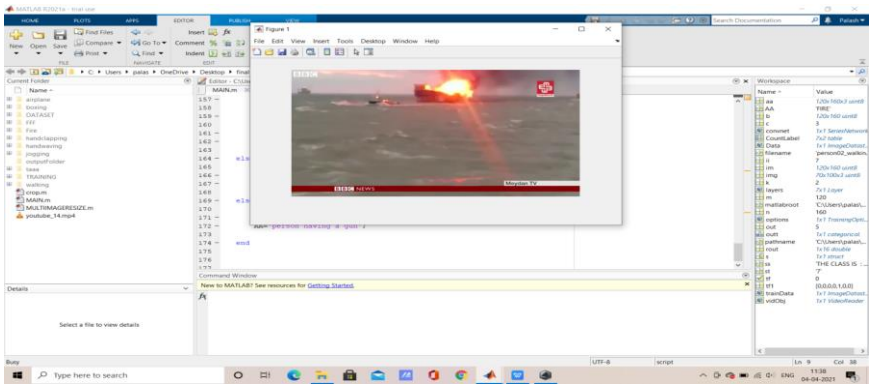

Figure 4.Image processing done using MATLAB image processing tool

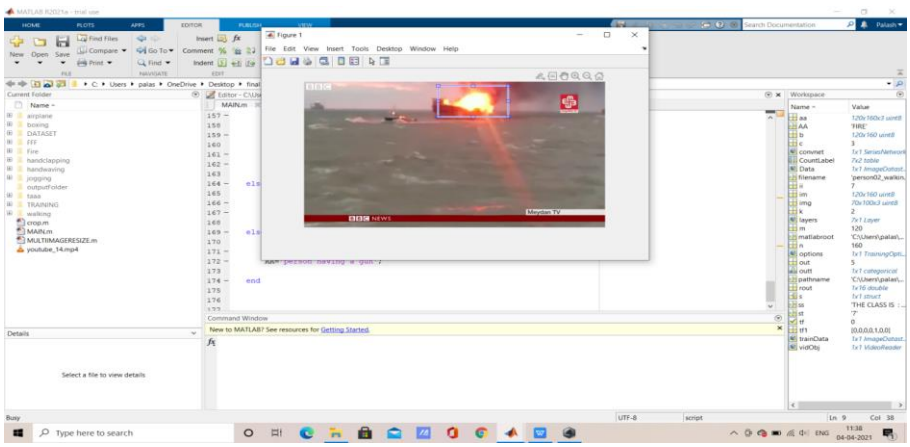

Figure 5. Deep Learning Tool analyzing the image

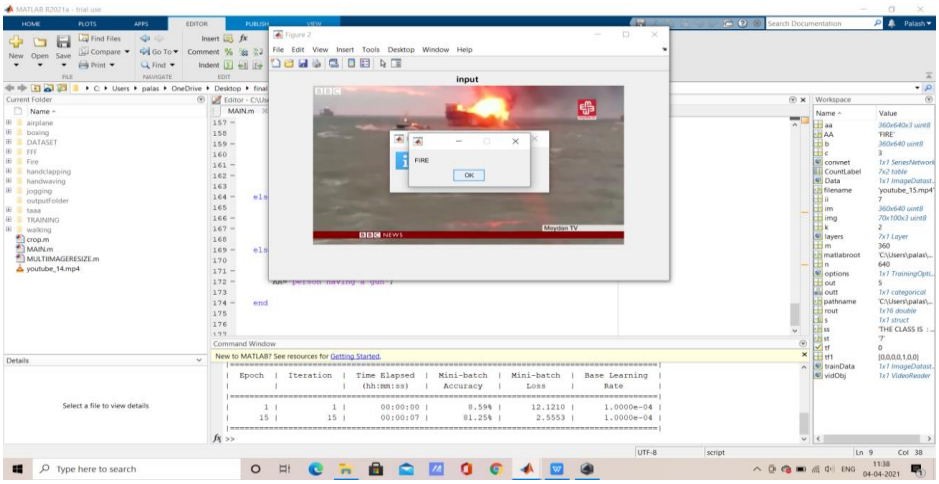

Figure 6. Trained Data

7. Conclusion

To accomplish great execution of video action recognition, we propose a classifier of CNN, which can acquire the information adjusted from pictures dependent on the regular visual highlights. In the interim, it can completely use the heterogeneous highlights of unlabeled recordings to upgrade the exhibition of activity acknowledgment in recordings. In our investigations, we approve that the information gained from pictures can impact the acknowledgment exactness of recordings and that distinctive acknowledgment results are acquired by utilizing diverse obvious signs. Trial results show that the proposed CNN has better execution of video activity acknowledgment, contrasted with the best-in-class techniques. And the exhibition of CNN is promising when just not many named training recordings are accessible.

References

1. B. Ma, L r. Huang, J. Shen, and L. Shao, "Discriminative of the tracking using tensor pooling," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2015.2477879.
2. L. Liu, L. Shao, X. Li, and K. Lu, "Learning spatio-temporal representations for action recognition: Agenetic programming approach," *IEEE Trans. Cybern.*, vol. 46,no. 1, pp. 158–170, Jan. 2016.
3. A. Khan, D. Windridge, and J. Kittler, "Multilevel Chinese takeaway process and label-based processes for rule induction in the context of automated sports video annotation," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1910–1923, Oct. 2014.
4. H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. Brit. Mach. Vis. Conf.*, London, U.K., 2009, pp. 124.1–124.11.
5. L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal Laplacian pyramid coding for action recognition," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 817–827, Jun.2014.
6. M.-Y. Chen and A. Hauptmann, "MoSIFT: Recognizing human actions in surveillance videos," School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-09-161,2009.
7. M. Yu, L. Liu, and L. Shao, "Structure-preserving binary representations for RGB- D action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2015.2491925.
8. L. Shao, L. Liu, and M. Yu, "Kernelizedmultiview projection for robust action recognition," *Int. J. Comput. Vis.*, 2015, doi:10.1007/s11263-015-0861-6.
9. Baskar, M., Renuka Devi, R., Ramkumar, J. *et al.* Region Centric Minutiae Propagation Measure Orient Forgery Detection with Finger Print Analysis in Health Care Systems. *Neural Process Lett* (2021). Springer, January 2021. https://doi.org/10.1007/s11063-020-10407-4.

10. Arulananth, T.S., Balaji, L., Baskar, M. *et al.* PCA Based Dimensional Data Reduction and Segmentation for DICOM Images. *Neural Process Lett* (2020). November 2020. https://doi.org/10.1007/s11063-020-10391-9.

11. M. Bask**ar**, J. Ramkumar, Ayush Bharadwaj, Yattik Sihag, "Discounts and Profitability Analysis using Data Visualization Techniques". International Journal of Advanced Science and Technology, Vol. 29, No.06, pp: 2258 – 2270, ISSN: 2005-4238, May 2020.

12. M .Baskar, J. Ramkumar, V.Venkateswara Reddy, G.Naveen Reddy, "Cricket Match Outcome Prediction using Machine Learning Techniques", International Journal of Advanced Science and Technology, Vol. 29, No. 4, pp: 1863-1871, ISSN: 2005-4238, April 2020.

13. M .Baskar, J. Ramkumar,Ritik Rathore, Raghav Kabra, "A Deep Learning Based Approach for Automatic Detection of Bike Riders with No Helmet and Number Plate Recognition", International Journal of Advanced Science and Technology, Vol. 29, No. 4, pp: 1844-1854, ISSN: 2005-4238, April 2020.