Human Activity from Surveillance Camera Using Deep Learning

K Sahadevaiah[1], Chellaboyina Yaswanth[2],

[1]Professor, Computer Science and Engineering Dept, Jawaharlal Nehru Technological University, Kakinada, AP, India.
[2]Post Graduate Student, Master of Technology (IT), Jawaharlal Nehru Technological University, Kakinada, AP, India.

---***---

Abstract: In recent years, skeleton-based action recognition has drawn a lot of attention. Because deep learning can extract pertinent information and achieve high recognition accuracy, it has been widely used in picture recognition. Deep learning's ultimate goal is to give machines the same capabilities as the human brain for data analysis and learning by helping them recognise patterns and principles in test data. The goal of this project is to use Media Pipe and deep learning techniques to accomplish robust and accurate human action recognition. The Media Pipe offers pre-trained models that eliminate the need for ongoing training and are useful for accurately identifying important locations on hands, faces, and human bodies. By monitoring their changes over time, these salient points can be utilised as characteristics for action recognition. The convolution and pooling layers of the convolutional neural network get the important information, resulting in an effective action prediction. Based on the input, the model predicts 12 distinct actions. Python is used to implement the deep learning algorithms and media pipe for human action recognition.

Keywords: Human Activity Recognition, Deep Learning, media pipe, Computer Vision.

INTRODUCTION

With the rising of crime rates become an issue if they are not promptly recognised and the appropriate safety measures are not implemented. The majority of cities and metropolitan areas have deployed surveillance systems that continuously gather data. The enormous amount of surveillance data means that there is a greater likelihood of suspicious activity. However, because these jobs are too complex and resource-intensive for artificial intelligence to undertake, human monitoring is necessary to detect such behaviours. One method to simplify an activity for automation is to break it down into smaller components and identify subtasks that could lead to potential crimes. We use our models to try and identify two primary pathways that could lead to crimes.

OBJECTIVE:

The primary goal is to use deep learning to construct the model for human action recognition. Probably demonstrating that artificial shallow neural networks are not the optimal method for classification; instead, deep learning.

In the modern world, the most fundamental & effective security measure a building can have is CCTV surveillance. Hospitals, shopping centres, universities, and other establishments use it as the most well-known means of identifying and stopping undesired activity. However, picture an academic campus with over 100 CCTV cameras spread over several structures, such as dorms, classrooms, canteens, sports areas, auditoriums, etc. It is not possible to manually watch every incident captured by the CCTV camera. It takes a lot of time to manually look for the identical incident in the recorded video, even if it has already happened. All things considered, we plan to create a single deep learning model that uses media pipe module data to forecast human behaviour. Lastly, we are contrasting the accuracy using the current ANN technique system.

DOMAIN OVERVIEW:

These days, deep neural networks are extensively used in both academia and business as the cutting edge of machine learning models in a range of fields, including natural language processing and image analysis.

Slowly but surely, these advancements hold great promise for medical imaging technologies, medical data analysis, medical diagnostics, and healthcare overall. We give a brief summary of current developments in machine learning as they relate to medical image processing and analysis, along with some related difficulties. Conventional machine learning techniques were the norm long before deep learning was employed. Like SVM, Logistic Regression, Decision Trees, and Naive Bayes Classifiers.

Another name for these algorithms is flat algorithms. Here, "flat" refers to the fact that these techniques are typically not able to be applied directly to the raw data (text, images,.csv files, etc.). A preprocessing procedure known as feature extraction is required.

These traditional machine learning algorithms can now employ the representation of the provided raw data as the outcome of feature extraction to complete a task. As an illustration, consider the division of the data into multiple classes or categories.

Feature extraction typically involves a great deal of complexity and in-depth understanding of the issue domain. For best results, this pre-processing layer needs to be adjusted, tested, and improved across a number of rounds.

DL based ANN are on the opposite side. The Feature Extraction phase is not necessary for these. The layers have the ability to independently and directly learn an implicit representation of the raw data. Here, multiple layers of artificial neural networks are used to create an increasingly abstract and compressed representation of the original input. The outcome is then generated using this condensed representation of the input data. One possible outcome is the input data being divided into various classes.

LITERATURE SURVEY

The most crucial stage in the software development process is the literature review. Determining the time factor, economics, and company strength is required before building the tool. The next stage is to decide which operating system and language can be used for tool development once these requirements are met. The programmers require a great deal of outside assistance once they begin developing the tool. This support can be obtained from senior programmers, from book or from websites. The aforementioned factors are taken into account before constructing the suggested system.

A literature review is a corpus of work that tries to review the important aspects of current knowledge, such as significant discoveries and theoretical and methodological advancements related to a specific subject. Since literature reviews are secondary sources, they don't present any brand-new or unique experimental research. A literature review can also be thought of as an assessment of an intangible achievement.

PROPOSED SYSTEM:

- Input video
- Pre-processing
- Feature extraction
- CNN (convolutional neural network)

ADVANTAGE:

- High accuracy
- It will identify the suspect with videos

Implementation Steps:

1. Import the necessary modules in Python.

2. Loading the dataset of human action photos and utilising the media pipe module for analysis.

3. Data augmentation, which divides the visual data into train and test sets, increases the appropriate information.

4. Create the framework needed to save the model and train the dataset.

5. Determining the maximum accuracy model and carrying out the validation process with test data.

6. Using the sample image information and the model to identify the human action.

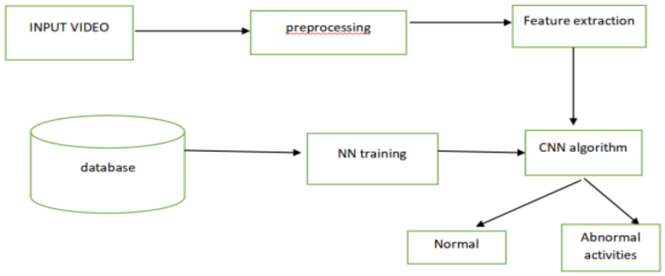

Figure 1: Block Diagram

Load Data: Extracting 15 labelled human activity photos from a collection of 15,000 images. It will be easier to extract the features if media outlets use their channels to clearly show human land marks.

Prior to processing: The processes done to format images before to their usage in model training and inference are known as image pre-processing. This covers resizing, aligning, and colour adjustments, among other things.

Practice and assess: Two subsets are created from the datasets. The first subset, known as the training data, contains a piece of our real dataset, which is used to help the machine learning model recognise and learn patterns from it. In this way, it trains our model. The other subset is the testing data.

Feature analysis:- According to feature analysis, we may see distinct qualities in every object and pattern we come across. According to the recognition-by-components theory, we identify objects by breaking them down into their constituent elements. These elements are recognised as geons, which are three-dimensional forms. Compared to directly applying machine learning to the raw data, it produces better outcomes.

CONVOLUTION NEURAL NETWORK (CNN)

CNN CNN though it may sound like a strange mash-up of math and biology with a dash of computer science, these networks have been among the most important developments in computer vision. Neural nets gained popularity for the first time in 2012 when Alex Krizhevsky used them to win the ImageNet competition, which is akin to the yearly Olympics of computer vision. At the time, this achievement was remarkable since it reduced the classification error record from 26% to 15%. Since then, deep learning has been the foundation of many businesses' offerings. Neural nets are used by Facebook for its automatic tagging algorithms, by Google for photo search, by Amazon for product suggestions, by Pinterest for personalised home feeds, and by Instagram for search infrastructure

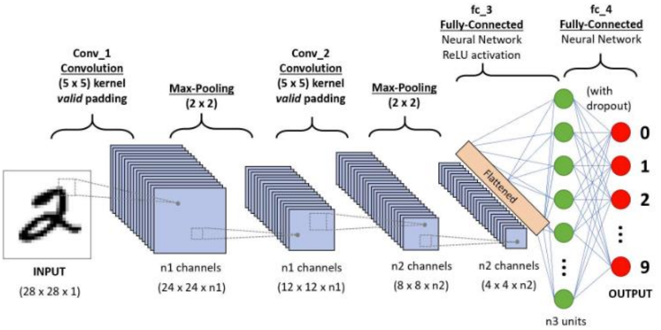

Figure 2: CNN Architecture

The Problem Space

The process of taking an input image and producing a class (a dog, cat, etc.) or a probability of classes that best describes the image is known as image classification. One of the first abilities we acquire as humans is the ability to recognise, which is a skill that comes easily and readily to adults. We can rapidly and easily recognise the objects around us and the surroundings we are in without giving it a second thought. Most of the time, without even realising it, we are able to describe a scene and assign a label to each thing when we see an image or simply when we gaze at the world around us. We do not share these abilities with other machines: the ability to swiftly identify patterns, generalise from past knowledge, and adapt to various visual settings.

Inputs and Outputs

When When a computer views (or receives) an image, it perceives an array of pixel values. The display of a 32 × 32 x 3 array of integers will depend on the size and resolution of the image (the 3 stands for RGB values). To emphasise the concept, let's take an example where we have a 480 × 480 colour JPG image. The 480 × 480 x 3 representative array will be used. A value between 0 and 255 is assigned to each of these values, denoting the pixel intensity at that particular location. The only inputs available to the computer for image classification are these numbers, even if they have no meaning for us.

What We Want the Computer to Do

Now that we are aware of the issue and its sources and results, let's consider our options for solving it. The computer is supposed to be able to distinguish between each image it is shown and identify the distinctive characteristics that define a dog as a dog or a cat as a cat. This is also the process that occurs subliminally in our minds. If an image of a dog contains distinguishable characteristics, like four legs or paws, we can categorise it as such. Similar to this, the computer can classify images by first searching for basic features like edges and curves, and then using a sequence of convolutional layers to advance to more complex ideas. This is a broad summary of what a CNN performs. Now let's talk about the details.

Biological Connection

First, though, some background. You might have associated CNN with neuroscience or biology when you first heard the word, and you would be correct. Kind of. CNNs are inspired biologically by the visual cortex. Different regions of the visual field elicit distinct responses from small cell areas in the visual cortex. Hubel and Wiesel's seminal 1962 experiment (Video) advanced this hypothesis further by showing that certain individual brain neuronal cells only fired, or responded, in the presence of edges oriented in a specific way. For example, certain neurons responded when they observed a vertical edge, while other neurons fired when they observed a horizontal or diagonal edge. Researchers Hubel and Wiesel found that these neurons may work together to create visual perception. The arrangement of these neurons was columnar. CNNs are predicated on the idea that specific components of a system have discrete roles (for example, the neurons in the visual cortex searching for particular traits). This concept is also used by computers.

Structure

Now let's return to the specifics. To provide a more comprehensive understanding, consider the following: a CNN processes an image by applying various fully connected, convolutional, nonlinear, pooling (down sampling), and other layers before producing the desired result. As we mentioned earlier, the result can be a single class or a probability of classes that best describe the image. Right now, figuring out what each of these levels does is the difficult part. Let's talk about the most important one now.

First Layer – Math Part

A CNN's convolutional layer is always its initial layer. The first thing you should keep in mind is what this conv (I'll be using that acronym a lot) layer's input is. As previously stated, the input consists of a 32 × 32 x 3 array of pixel values. Currently, the best approach to visualise a convolution layer is to see the upper left portion of the image illuminated by a torch. Assume that a 5 5 area is illuminated by the light this torch emits. Let's now visualise this torch moving across every section of the input image. This torch is known as a filter in machine learning terminology (it is also sometimes termed a neuron or kernel). The area it illuminates is known as the receptive field. Currently, this filter also consists of an array of numbers, which are referred to as weights or parameters. The dimensions of this filter are 5 x 5 x 3, which is a very significant detail since the depth of the filter must match the depth of the input in order for the math to work. Let's now use the filter's initial location as an example. That is the upper left corner. The filter multiplies its values by the original pixel values of the input image while it slides, or convolves, around it (also known as computing element-wise multiplications). All of these multiplications are added together, or 75 multiplications overall in mathematics. You are now down to one number. Recall that this figure merely indicates the moment in the picture when the filter is at the upper left corner. We now carry out this procedure one again for each place on the input volume. The filter would then be moved one unit to the right, one unit to the right again, and so forth. A number is produced at each distinct position on the input volume. All that's left when you slide the filter over each area is an array of integers measuring 28 by 28 by 1, which is referred to as an activation map or feature map. A 5 × 5 filter can fit in 784 distinct spots on a 32 x 32 input image, which is why you get a 28 x 28 array. A 28 × 28 array is mapped to these 784 numbers.

First Layer – High Level Perspective

Let's take a high-level look at what this convolution is truly doing, though. You can think of each of these filters as a feature identifier. When I refer to features, I mean characteristics like curves, plain colours, and straight edges. Consider the most basic qualities that all photographs share with one another. Assume that our first filter, a 7 x 7 x 3 curve detector, will function as such. (For the sake of simplicity, let's ignore the filter's three units of depth in this part and just focus on the image's top depth slice.) The filter functions as a curve detector by having a pixel structure with larger numerical values along the areas that form curves (keep in mind that these filters are just numbers!).

Convolution Neural Network Creation Steps

1. Specify the Input Layer: To begin, specify the input layer's dimensions (such as the image's height, width, and number of channels).

2. Incorporate Convolution Layers:

Use a convolutional layer or layers to extract features from the input data. For every layer, specify characteristics such the activation function, kernel size, and number of filters.

3. use Pooling Layers: To downsample the feature maps produced by the convolutional layers, use pooling layers (such as max pooling). Pooling facilitates the reduction of spatial dimensions and the management of overfitting.

4. Flatten the Output: Convolutional or pooling layer's final output should be flattened into a 1D vector. In order to feed the data into the fully connected layers, this prepares it.

5. Add Fully Connected Layers: To process the flattened characteristics, add one or more fully connected (dense) layers. Indicate the activation functions and the quantity of units (neurons) for each of these layers.

6. Define the Output Layer: Depending on the type of task you're doing (binary classification, multi-class classification, etc.), choose an appropriate activation function (like SoftMax for classification) in the output layer.

7. Assemble the Model: Assemble the CNN model by indicating the optimizer, loss function, and metrics that will be applied to training.

8. Train the Model: A training dataset is used to train the CNN model. To maximise performance, modify hyperparameters (such as batch size and learning rate) as necessary.

9. Assess the Model: To determine the model's performance (e.g., accuracy, loss), evaluate the trained model using a different validation dataset.

Forecasts: Apply the learned CNN model to forecast fresh or unobserved data. Consider the task when interpreting the model outputs (e.g., class probabilities for classification).

Parameters Calculations

Initially, the confusion matrix was used to assess the outcomes of binomial classification. Therefore, the first step is to designate the positive class—one of the two classes—as the class of interest. In the target column, we have to select (randomly) one value to represent the affirmative class. The other value is then automatically allocated the negative class. Although this assignment is entirely random, keep in mind that some class statistics will show different values based on the selected positive class. In this example, we utilised the regular e as the negative class and the positive class.

True Positive(TP)	False Positive(FP)
False Negative(FN)	True Negative(TN)

$$\text{Accuracy (ACC)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Specificity (Spc)} = \frac{\text{No of TN}}{\text{No of TN} + \text{No of FP}}$$

$$\text{Recall (Re)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1} - \text{Score (F1S)} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

Results:

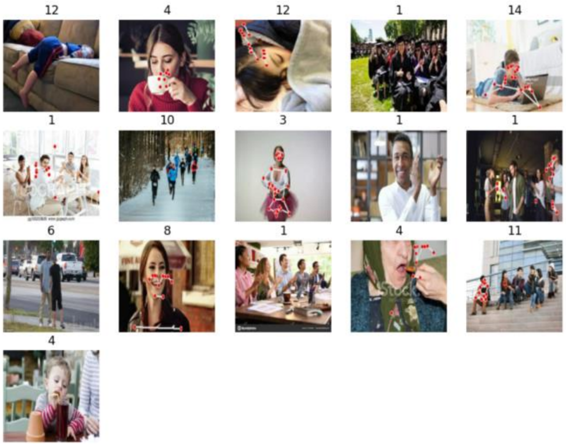

Figure 3: Media pipe visualization result of the sample data

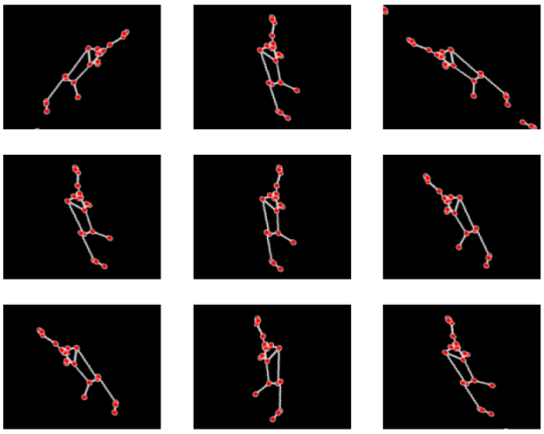

Figure 4: Data augmentation media pipe result for the input data

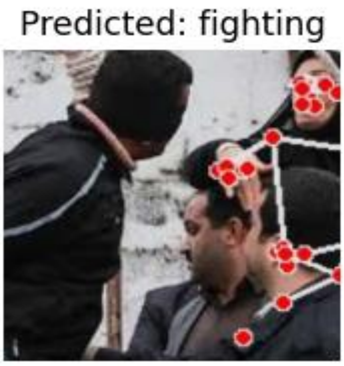

Figure 5: Final result of model label prediction for the input image.

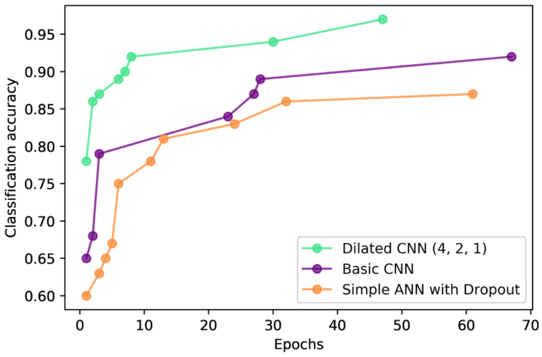

Figure 6: Comparison result for the existing system and our proposed system

Note that this is a experimental result for comparison of CNN models with ANN model, where we are implementing the ANN base model in our project.

CONCLUSION

Using Media Pipe and Deep Learning, a Human Action Recognition system was successfully constructed for this project. A dataset including 20,000 annotated skeletal photos of 15 distinct human activities was used to train the machine. The ResNet-18 architectural model is used, which was found to provide effective training. The model proved its efficacy in identifying human motions from skeletal photos by achieving accuracy on the testing set. MediaPipe was used to derive the skeleton pictures from the original photographs, which supplied essential position information for action recognition. Applications in the actual world, including action recognition in surveillance, are possible. In the end, our project's use of convolution neural networks allows us to create models with up to 85% accuracy.

FUTURE SCOPE:

The potential uses of DL with surveillance cameras for monitoring human behaviour appear to be extensive and span a wide range of fields. To guarantee the proper deployment and use of these technologies, it is imperative to address privacy, security, and ethical issues.

REFERENCE:

[1] S. Ren, K. He, R. Girshick, and J. Sun, "[Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks](https://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks)," in NIPS 2015 - Advances in Neural Information Processing Systems 28. Neural Information Processing Systems Foundation, 2015

[2] Pierre Sermanet David Eigen Xiang Zhang Michael Mathieu Rob Fergus Yann LeCun "[OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks](https://arxiv.org/pdf/1312.6229.pdf),"

Courant Institute of Mathematical Sciences, New York University.

[3] The IMFDB Internet Movie Firearms Database, [Online] Available: http://www.imfdb.org/wiki/Main_Page [Accessed Mar 20, 2019]

[4] Soft Computing and Intelligence Information Systems, [Online] Available: https://sci2s.ugr.es/weapons-detection [Accessed Mar 27, 2019]

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "[Rethinking the Inception Architecture for Computer Vision](https://arxiv.org/abs/1512.00567)," in arXiv, vol. abs/1512.00567, 2015.

[6] Rohit Kumar Tiwari and Gyanendra K. Verma, "[A Computer Vision based Framework for Visual Gun Detection using Harris Interest Point Detector](https://www.sciencedirect.com/science/article/pii/S1877050915002010)," Procedia Computer Science, vol 54, p. 703 - 712, 2015.

[7] Samir K. Bandyopadhyay, Biswajita Datta, and Sudipta Roy "[Identifications of concealed weapon in a Human Body](https://www.sciencedirect.com/science/article/pii/S1877050911020673)," Department of Computer Science and Engineer, University of Calcutta, 2012

[8] Claire-HlneDemarty, et. al "[The MediaEval 2012 affect task: violent scenes detection](https://hal.inria.fr/hal-00731568/document)," Working Notes Proceedings of the MediaEval 2012 Workshop. 2012.

[9] Roberto Olmos, Siham Tabik, and Francisco Herrera "[Automatic Handgun Detection Alarm in Videos Using Deep Learning](https://www.mdpi.com/2076-3417/7/4/385)," Soft Computing and Intelligent Information Systems research group, Department of Computer Science and Artificial Intelligence, University of Granada, 2017

[10] Lisa Torrey and Jude Shavlik, "[Transfer Learning](http://ftp.cs.wisc.edu/machine-learning/shavlikgroup/torrey.handbook09.pdf)," University of Wisconsin, Madison WI, USA.

About Authors

Mr. Chellaboyina Yaswanth, Post Graduate Student (M.Tech -IT), Computer Science and Engineering Dept, Jawaharlal Nehru Technological University, Kakinada. His Area of Interest includes Internet of Things, Deep Learning, and Computer Networks.

Dr.K.Sahadevaiah, Professor, Computer Science and Engineering Dept, Jawaharlal Nehru Technological University, Kakinada. His areas of interest include Ad Hoc Wireless Networks, Vehicular Ad Hoc Networks, Wireless Sensor Networks, and Wireless Mesh Networks.