Survey on Human Behavior Recognition using CNN

Anushree Raj[1], Sadiya Ayub Humbarkar[2], Sumedha E[3]

[1] *Assistant Professor- IT Department, AIMIT, Mangaluru, anushreeraj@staloysius.ac.in*
[2] *MCA Student, AIMIT, Mangaluru, 2117097SADIYA@staloysius.ac.in*
[3] *MCA Student, AIMIT, Mangaluru, 2117112SUMEDHA@staloysius.ac.in*

---***---

Abstract– Human behavior recognition is a crucial area of scientific research in the science of computer vision that has significant applications in a variety of industries, including intelligent surveillance, smart homes, and virtual reality. Traditional manual approaches have a hard time meeting the demands of high recognition accuracy and applicability in the contemporary complicated environment. Deep learning's arrival has opened up new avenues for behavior recognition research. The major focus of this paper is behavior recognition using convolutional neural networks (CNN). Before discussing and analyzing the classical learning methods and deep learning methods of behavior recognition, the research context and importance of behavior recognition are first introduced. Based on the convolution neural network designed for the specific human behavior in public areas, we develop a series of human behavior recognition systems. In order to extract moving foreground characters of the body, the video of human behavior data set will first be divided into images, which will then be processed using the background removal approach. Second, the planned convolution neural network is trained with the training data sets, and the depth learning network is built using stochastic gradient descent. Finally, using the developed network model, the numerous sample behaviors are categorized and recognized, and the recognition outcomes are evaluated against the state-of-the-art techniques. The findings demonstrate that CNN is capable of studying human behavior models automatically and recognizing human behaviors without the need for manually annotated training.

Keywords—Convolutional Neural Network (CNN); deep learning; YOLOv3 algorithm; LSTM (Long Short-Term Memory networks); R-CNN (Region-Based Convolutional Neural Network)

1. INTRODUCTION

The technique of classifying and recognizing human behaviors, such as activities or expressions, is based on observations. The global aspects of a picture have become increasingly important in traditional human behavior identification over the past few decades. To characterize human behavior, these static elements include edge features, shape features, statistical features, and transform features. It is a method for categorizing and detecting activities based on observations, like sensor data streams. Recognition of Human Behavior involves several processes, including detection, description, clustering, and recognition. Recently, the field of ubiquitous computing has made major advancements in the study of device-free human behavior recognition as well as behavior recognition in video and photos.

There are a number of techniques for recognizing human activity, including conventional techniques that rely on features retrieved from photos. Deep learning is the most advanced technology for object detection, processing and detecting vast amounts of picture data with the least amount of latency. The CNN model-based behavior recognition implementation has received a lot of attention.

Convolutional Neural Network (CNN) is a network with a focus on computer vision, a class of deep learning models that are mostly used for object detection and picture processing. A lightweight convolutional neural network is created for the purpose of recognizing human behavior in order to lower the number of network parameters and lower the demand for processing and storage resources. It is suggested to use a combined training approach that combines pre-training, fine-tuning training, and migration training to increase the deep CNN model's performance at recognition. When we input an image into a CNN, it has numerous layers, and as each layer generates activation functions, the following layer receives them. The network may recognize increasingly more complex elements, including objects, faces, etc., as we go further into it.

2. OBJECTIVE

The objective of this paper is to show the capability of deep learning to implement the human behavior recognition. To regulate and achieve the recognition of human behaviors activities in real time, a Convolutional Neural Network (or CNN) framework is established. The augmentation is considered for training data set expecting better prediction. The prime task of this research paper is to take out visual information from digital video; where human body movement needs to be captured. A myriad of digital video and image processing techniques is suitable for extracting information, factual characteristics of human behavior observation. It requires the technologies such as digital image processing, pattern recognition, machine learning, and deep learning.

3. RELATED WORKS

To recognize pedestrians, Jia Lu, Wei Qi Yan, and Minh Nguyen demonstrated a deep learning-based detection method. The study used the YOLO model, a deep learning technique that enables real-time detection. To reduce the time-consuming, a GPU acceleration is needed while deep learning is being trained and tested. A suitable hyperparameter should be carefully chosen in order to fine-tune the model because various hyperparameters can influence the outcomes. Extending the YOLO detection approach should be the focus of future development. Deep learning demonstrates the capacity to recognize objects and assign each one to the appropriate class.

Mayur Shitole, Jerry Zeyu Gao, Shuqin Wang, and Hanping Lin Sheng Zhou and Layla Reza propose well-defined emoji-based human behavior patterns to facilitate machine learning-based dynamic behavior classification and detection. It concentrates on four different human actions: standing, moving quickly, moving slowly, and sitting. Additionally, a system is described to facilitate real-time human dynamic behavior identification and categorization based on the suggested machine learning model and emoji-based behavior patterns. The paper also presents some prior case study results for dynamic human behavior detection and classification utilizing emoji representation. Live streaming as well as pre-recorded videos can both be played on the system without any issues.

A deep network and HMM-based behavior recognition technique is proposed by Chen Chen. This study maximizes the benefits of traditional approaches to retain features by combining them with deep learning techniques. The benefits of deep networks, self-extraction, self-training, and time information processing allow his suggested strategy to have a positive impact on the identification of interactive behavior. However, this method is still not timely due to the laborious manual extraction of features by conventional methods.

Zhengjie Wang and Yinjing Guo provide the current general methods of behavior identification, along with related surveys, the concept of channel state information, and an explanation of the principles of CSI-based behavior recognition. The paper also goes into great detail about the general framework for behavior recognition, including the fundamental signal selection, signal preprocessing, and behavior identification techniques employing pattern-based, model-based, and deep learning-based approaches. The paper divides the existing research and applications into three categories based on the aforementioned recognition methodologies and describes each typical application in detail, including the test equipment, experimental situations, user count, observed behaviors, classifier, and system performance. Additionally, it examines a few particular applications and includes in-depth discussions on the choice of recognition methods and performance assessment. These conversations offer some valuable suggestions for creating an identifying system.

Not much thought was given to which way the AcFR system would move when it decided to alter its perspective. This can be problematic since the system might choose to examine the person from behind rather than the front, which is how individuals typically move to see a subject's face more clearly. For more accurate active face recognition, the direction of the face must be estimated.

Using EEG brain waves, Sumin Jin, Yungcheol Byun, and Sangyong Byun have suggested a method to identify specific human behaviors or actions. They identified six behaviors for this and recorded the brain waves associated with each behavior. They used CNN and LSTM models, and the studies revealed that they were able to recognize 66% of behaviors using EEG brain waves. This is a promising result given the complexity of the interaction between brain waves and behaviors. Due to dynamic information, the LSTM model produced a better outcome.

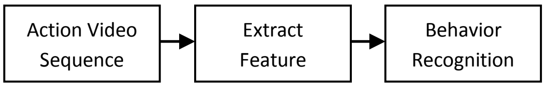
An enhanced deep learning-based method for identifying abnormal human behavior is proposed by Weihu Zhang and Chang Liu. This approach has a greater rate of recognition, extracts feature more precisely, and simplifies the model less than the conventional approach. To obtain precise critical areas of human motion and an optical flow map, the Gauss model is chosen, and the Farneback dense optical flow technique is applied. The benefits of CNN and LSTM are combined to produce an accurate recognition effect.

Franjo Matkovi, Darijan Mareti, and Slobodan Ribari offer a method for identifying motion patterns and abnormal crowd behavior in surveillance film. It is based on an analysis of fuzzy predicates and fuzzy logic formulas derived from human interpretation of real video sequences, (multi-agent) crowd simulators, and data from common sense. To identify and categorize motion patterns in line with the given taxonomy of fuzzy logic predicates, fuzzy logic predicates are used to analyze the motion patterns of an individual or group of individuals. The detection and classification of unusual crowd behavior using fuzzy logic functions. The fuzzy predicates serve as the fundamental building blocks of fuzzy logic functions, and the assignment functions for these predicates are created by expertly interpreting training video sequences in conjunction with fuzzy logic operators. We use genuine trajectories obtained by the proposed 4-pipelined multi-person tracker and ground truth annotations of actual video sequences to evaluate the proposed technique. Positive and reassuring results have been found in early tests.

To address the issue of long-term modelling of existing behavior recognition algorithms, Feng Xiufang and Dong Xiaoyu suggested a group feature behavior recognition algorithm based on the attention mechanism. The redundant frames between frames in video sequences are successfully eliminated using sparse sampling. The original frame images are used to model space features in CNN to efficiently extract motion change information. Progressive networks with pyramid pools are used to extract picture features during network training. The final feature vector is then produced by adding an attention layer after the video frame features have been consecutively encoded by Bi-GRU. The experimental results demonstrate that this paper's data features can effectively enhance the network's ability to express itself, and that this paper's network structure can well mimic the long-term attention of videos.

4. METHODOLOGY

As illustrated here, the identification and comprehension of human behavior feature extraction and motion are the two fundamental components of human behavior recognition. The process of feature extraction involves taking the important features from video or picture data. Since feature information is crucial for recognition, feature extraction really has a direct bearing on the outcome of recognition.



Data Collection

There are numerous publicly available datasets for the identification of human behavior, including the Weizmann dataset, UT-Interaction dataset, KTH dataset, UCF dataset, BEHAVE dataset, HMDB51 dataset, and MS COCO dataset. A brief summary of these datasets is included in the table 1 below.

Datasets	***Brief Description***
Weizmann	consists of 90 movies of nine participants doing 10 distinct movements, including sprinting, jumping in place, forward jumping, bending, waving one hand, jumping jacks, side jumping, standing on one leg, strolling, and waving two hands.
UT-Interaction	Contains footage of human-to-human encounters from the six classes of handshake, point, hug, push, kick, and punch performed continuously.
KTH	contains the following six actions: hand clap, box, jog, walk, and jog. Each action is performed by 25 different people, and the setting is systematically changed for each actor's action to accommodate for performance nuance.
UCF	contains 13,320 video clips that are divided into 101 different categories. There are 5 types that can be assigned to these 101 categories (Body motion, Human-human interactions, Human-object interactions, Playing musical instruments and Sports). The videos have all been compiled from YouTube.
BEHAVE	a collection of data on interactions between people and objects in the wild. It features 20 objects being used by 8 persons in 5 different natural settings.
HMDB51	a compilation of realistic footage taken from a range of media, including television and the web. 6,849 video clips from 51 action categories make up the collection.
MS COCO	a large collection of 328,000 pictures of people and common objects.

Table 1: Summary of these datasets

Data Pre-Processing

The raw data gathered by motion sensors needs to be pre-processed in the following ways in order to feed the suggested network with a certain data dimension and increase the model's accuracy.

1) Linear Interpolation: The aforementioned datasets are accurate, and the subjects wore wireless sensors. As a result, throughout the collection procedure, some data could be lost. NAN/0 is commonly used to indicate the missing data in these circumstances. This problem was resolved by employing the linear interpolation method to fill in the missing variables in this investigation.

2) Scaling and Normalization: It is essential to normalize the input data to the 0–1 range because training models straight from large values from channels may fail.

Data Augmentation

A large scale of dataset is the premise of a successful application of convolutional neural networks (CNNs). In order to create training samples and increase the size of the training dataset, data augmentation methods alter the training image in a number of random ways. Increasing the depth and width of a neural network typically improves its learning capacity, making it easier to fit the distribution of training data. Our research demonstrates that in the convolution neural network, depth is more significant than width. However, as the depth of neural networks increases, so do the parameters that must be taught, which will result in overfitting. Too many parameters will fit the properties of the dataset when it is small. The data augmentation consists of random noises, scale, rotation, and crop.

Proposed Method

We suggest a YOLOv3 strategy to identifying and classifying dynamic human behavior patterns, which is motivated by prior research and methodologies. A real-time object detection system called YOLOv3 (You Only Look Once, Version 3) recognizes particular things in films, live feeds, or still photos. To find an item, the YOLO machine learning system leverages features that a deep convolutional neural network has learned.

A Convolutional Neural Network (CNN) called YOLO is capable of quickly recognising things. CNNs are classifier-based systems that are able to examine input images as organised arrays of data and find connections between them. YOLO has the advantage of being faster than other networks while keeping accuracy. The model can now see the complete image at test time, which helps it make more accurate predictions. Regions are scored by YOLO and other convolutional neural network algorithms based on how closely they resemble predetermined classifications.

Initially, the YOLOv3 algorithm divides an image into a grid. Each grid cell foretells the presence of a specific number of boundary boxes (also known as anchor boxes) around items that perform well in the aforementioned predetermined classes. Only one object is detected by each boundary box, which has a corresponding confidence score indicating how correct it expects that prediction to be. The ground truth boxes' dimensions from the original dataset are clustered to find the most prevalent sizes and shapes before creating the border boxes.

R-CNN (Region-based Convolutional Neural Networks, developed in 2015), Fast R-CNN (an R-CNN upgrade developed in 2017), and Mask R-CNN are further comparable algorithms that can accomplish the same goal. However, YOLO is taught to do classification and bounding box regression simultaneously, in contrast to systems like R-CNN and Fast R-CNN.

5. ANALYSIS

Compared to past studies of behavior recognition using CNNs, this study gives more understandings with respect to accuracy on human behaviors. With the use of this study approach, it is possible to anticipate human behavior more accurately from raw data while also simplifying the model and doing away with the necessity for sophisticated feature engineering. Selection of Datasets are decided on the basis of accuracy and complexities and its features that can be obtained from it. This study can be further extended and can be implemented for different behaviors. Although many implemented this recognition method using different models like LSTM, YOLO, R-CNN models, most important features are obtained from CNN model.

6. CONCLUSION

In this research study, human behavior and activity is recognized using convolutional neural network. Human behavior identification is a complex task that is why series of images have been analyzed so that every moment can be captured for analysis and prediction. For more information as a dataset is prepared by data augmentation process. More training data makes the system robust and more accurate. Deep learning process includes multidimensional input data set and has the capability of heretical, sequential calculation simultaneously along with adaptation process. This ability makes it suitable for behavior analysis. To make understand the networks video clip converted into series of images so that machine can learn deeply every moment of human activity. In comparison to other approaches, the suggested approach can efficiently cut down on complexity while maintaining network performance. Future research can develop a new parameter selection technique to boost recognition performance even more. Deep learning network has the scope of weight updating which is useful for dynamic behavior identification that is the resultant of behavior change activity.

REFERENCES

[1] Jia Lu, Wei Qi Yan and Minh Nguyen, "Human Behaviour Recognition Using Deep Learning", 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS).

[2] Shuqin Wang, Jerry Zeyu Gao, Hanping Lin, Mayur Shitole Layla Reza, Sheng Zhou, "Dynamic Human Behavior Pattern Detection and classification", 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService).

[3] An Gong, Chen Chen, and Mengtang Peng, "Human Interaction Recognition Based on Deep Learning and HMM", IEEE Access 2019(Volume: 7).

[4] Zhengjie Wang, Kangkang Jiang, Yushan Hou, Wenwen Dou, Chengming Zhang, Zehua Huang, and Yinjing Guo, "A Survey on Human Behavior Recognition Using Channel State Information", IEEE Access 2019 (Volume: 7).

[5] Chenxi Huang, Yutian Xiao, and Gaowei Xu, "Predicting Human Intention-Behavior Through EEG Signal Analysis Using Multi-Scale CNN", IEEE/ACM Transactions on Computational Biology and Bioinformatics, IEEE 2020, Volume: 18.

[6] Masaki Nakada, Han Wang, Demetri Terzopoulos,''AcFR: Active Face Recognition Using Convolutional Neural Networks", 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

[7] Sumin Jin, Yungcheol Byun, Sangyong Byun,"Analysis of Brain Waves for Detecting Behaviors" , 2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Volume: 3.

[8] Weihu Zhang; Chang Liu," Research on Human Abnormal Behavior Detection Based on Deep Learning", 2020 International Conference on Virtual Reality and Intelligent Systems (ICVRIS).

[9] Franjo Matkoviü, Darijan Marþetiü, Slobodan Ribariü, "Abnormal Crowd Behaviour Recognition in Surveillance Videos", 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS).

[10] Feng Xiufang, Dong Xiaoyu, "Research on Human Behavior Recognition Method Based on Static and Dynamic History Sequence", 2020 Eighth International Conference on Advanced Cloud and Big Data (CBD).