Human Activity Recognition using Deep Learning

[1]M. Jeyapradha,[2]A. Raja Soorya

Assistant Professor
Department of Computer Applications
[1]The Standard Fireworks Rajaratnam College for Women, Sivakasi,India

Abstract: Vision-based human activity recognition in privately owned spaces and public places has become a significant issue in terms of developing safety feature to avoid theft like crimes. This technology improves security of privately owned spaces and public places. Deep learning models aim to automate extraction of low-level to high-level features of input data using convolutional feature instead of using aged complicated extraction methods. Convolutional feature has achieved significant improvements in classification of large amount of data especially in vision-based datasets. In order to recognize human action in smart environment, DMLSmartActions dataset is used for monitoring human activities. In deep learning, convolutional neural networks (CNNs) architecture is a class of deep neural networks commonly used to analyze visual imagery. The performance of the proposed method has been compared with previous methods that have used traditional machine learning methods on the same dataset. Experimental results demonstrated that the proposed deep learning model has achieved 82.41% accuracy rate in the classification of human activity which is far better than traditional machine learning method.

Keywords: DMLSmart Actions, Human Activity Recognition, Convolutional Neural Network, Human Image Threshing, Facial Image Threshing.

I. Introduction

Human Activity Recognition (HAR) has been a challenging problem, yet it needs to be resolved. It will be mainly used for eldercare and healthcare systems as an assistive technology when ensemble with other technologies like Internet of Things (IoT). Deep learning methods have been achieving success on HAR problems, which has given their ability to automatically learn higher-order features.Deep learning is a subdivisionof machine learning that tells computer systems to do what comes naturally to humans like learn from experience. Deep learning utilizes neural networks to learn useful representations of features directly from various types of data. Neural networks combine multiple nonlinear processing layers using simple elements operating in parallel. Deep learning models can accomplish the accuracy state in object classification, which can alsobe exceeding human-level performance. Using Convolutional Neural Network(CNN) pattern algorithm, recognition approaches have made tremendous progress in the past years. In this regard, in order to perform data classification and detection, mostly vision-based traditional pattern of machine learning methods have been used for recognition. In traditional methods, in order to acquire the features of video frames or images, complicated handcraft methods had been used. Vision-based recognizing human activity is a complex method. Therefore, new technique of obtaining the features of video frames or images have been introduced as Convolutional Neural Network (CNN) for recognition.

II. Literature Review

Md. Milon Islam, Sheikh Nooruddin, Fakhri Karray and Ghulam Muhammad proposed a paper “Human Activity Recognition Using Tools of Convolutional Neural Networks: A State of the Art Review, Data Sets, Challenges and Future Prospects”. This paper is based on Human Activity Recognition, Convolutional Neural Network, Multimodal Sensing Devices, Smartphone Data, Radar Signal, Vision Systems.

Ms. Shikha, Rohan Kumar, Shivam Aggarwal, Shrey Jain proposed a paper “Human Activity Recognition”. In this paper, an intelligent human activity recognition system is developed. Convolutional neural network (CNN) with spatiotemporal three dimensional (3D) kernels are trained using Kinetics data set. the results show promising activity recognition of over 400 human actions.

Namrata Roy, Rafiul Ahmed, Mohammad Rezwanul Huq, Mohammad Munem Shahriar proposed a paper “User-centric Activity Recognition and Prediction Model using Machine Learning Algorithms”. This work proposes a conceptual model that uses machine learning algorithms to detect activity from sensor data and predict the next activity from the previously seen activity sequence. It also measured the performance of an LSTM sequence prediction model for predicting the next activity with accuracy 70.90%.

III. Proposed Work

CNNs are one of the most popular architectures of deep learning which simulate biological nervous system like Artificial Neural Networks (ANNs). AlexNet, GoogleNet, SqueezNet and ResNet are the most common architectures of CNN. In comparison with ANNs, CNNs take the advantage of local connections instead of full connections in all layers except the last layer. In this regard, each layer by using the kernels or filter banks connects to the local region of the previous layer. Moreover, CNNs structure contains series of most common layers: The first layer is convolutional layer. In this layer, each region that contains feature maps that have been connect to the feature maps of a local region in the previous layer by calculating weights that are known as kernels (filter banks). Sum of all local weights goes through a non-linearity function.

IV. Advantages of Proposed System

Extracting and using large scale datasets of deep learning methods, which achieved state of art results in terms of detecting and recognizing human activity using the architecture of CNNs. Instead of using commonplace CNNs, a special CNN architecture is used to recognize the human activity has been designed. Additionally, the performance of the proposed method has been compared with the other previous used methods on the same dataset. In this study in order to classify the human activity frames, instead of using common CNNs, architecture of a specific CNN has been proposed which contains five convolutional layers, four pooling layers and three fully connected layers. In the last fully connected layer, softmax is considered to determine the probability of the 12 classes of activity dataset.

V. METHODOLOGY

5.1 Dataset

In Human Activity Recognition, the dataset of the Digital Multimedia Lab has been used. In order to create this dataset, the real and daily actions of seventeen people are captured by three static cameras in two simulated living environments. Moreover, videos contain twelve different natural activities of people, i.e. using cellphone, walking, writing, reading, sitting down, standing up, putting something, picking something, dropping and picking up, drinking, cleaning table and falling down. Using the histogram of oriented gradient (HOG) features from each video frame can represent human action. HOGs are well recognized for human detection and are mostly independent regarding illumination and contrast changes.

The method is based on computing well-normalized local histograms of image gradient orientations in a dense grid. The fundamental belief is that local object shape and appearance can often represent quite well by the distribution of local intensity gradients, even without accurate knowledge of the corresponding gradient or edge positions. This is implemented by separating the image window into small spatial regions, and each cell accumulates a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell.

The collective histogram entries form the representation. For better invariance to illumination, and shadowing, it is also supportive to contrast-normalize the local responses before applying them. This can be done by accumulating a measure of local histogram energy over to some extent larger spatial regions (blocks) and using the results to normalize all the cells within the block. The normalized descriptor blocks are referred to as Histogram of Oriented Gradient (HOG) descriptors.

5.2 Convolutional Neural Network

Deep learning algorithms have become more popular because of their automatic feature extraction capability from vision or image data as well as time-series data that enable to learn highlevel and meaningful features easily. Deep learning techniques is the outperforming traditional machine learning approaches for activity recognition in terms of classification performance measures such as accuracy, precision, recall, and F1 Score.

The recognition of human behavior based on deep learning architecture, CNN is categorized into various key stages. Overall system architecture with the following stages is illustrated in Figure 1 to understand CNN-based activity recognition. The first step is to select and implement the input devices like sensors, and cameras. Data collection is the next step where an edge device is used to perceive data from input devices and transfers it to the main server through various communication protocols like Wi-Fi, and Bluetooth. The deployment of computing and storage resources at the point where data being collected and processed is referred to as edge computing that incorporates sensors for data perception as well as edge servers for reliable real-time information processing. The feature extraction and selection stage extracts the necessary features from the raw signals; this stage is performed automatically in the case of CNN; no hand-crafted feature extractions are required. This stage contains the CNN architecture or variants of CNN structure for the recognition of activities. The last step includes a notification system through which an agent (human or machine) can be notified.

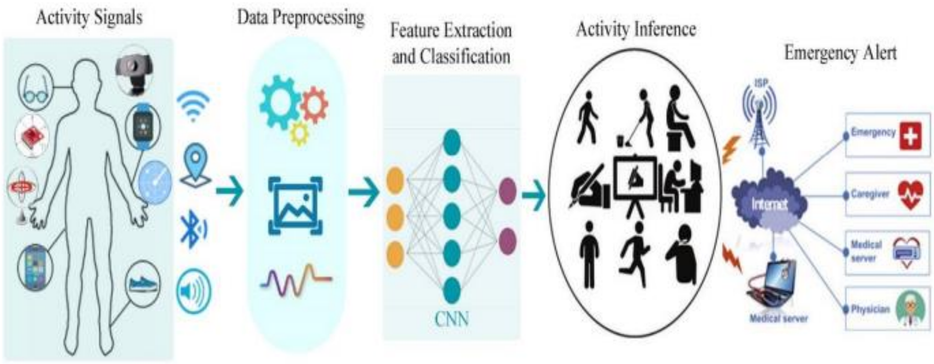

Figure 1 - Overall system architecture of CNN-based human activity recognition.

6.3 Evaluation Metrics

The multi-class confusion matrix is used to investigate the performance of the classification methods of multi-class datasets. Multi-class confusion matrix contains two dimensions which indicate the predicted classes by classifier and the actual classes. Non-negative sparse coding is a process for decomposing multivariate data into non-negative sparse components. In this, data representation and its relation to standard sparse coding and non-negative matrix factorization has been briefly described. Simple efficient multiplicative algorithm for finding the optimal values of the hidden components is given. In addition, how the basis vectors can be learned from the observed data has shown. Simulations demonstrate the effectiveness of the proposed method.

6.4 Image Analysis

The fundamental problem of image analysis is pattern recognition corresponding to physical objects in the picture, and determines their pose. Often the results of pattern recognition are all that’s needed, for example a robot guidance system supplies an object’s pose to a robot, and in other cases a pattern recognition step is required to find an object so that it can be inspected for defects.

GPM can replace NC template matching as the method for industrial pattern recognition. Template methods suffer from essential boundaries imposed by the pixel grid nature of the template itself. Translating, rotating and sizing grids by non-integer amounts require re-sampling, which is time consuming and limited accuracy. This confines the pose accuracy that can be achieved with template-based pattern recognition. Pixel grids and other represent patterns using gray-scale shading, which has been observed is often not reliable.

Sparse Coding is most likely one of the decomposition/factorization methods try to represent a signal with different atoms. In sparse coding, atoms are sparse and absolute. Sparse coding methods are used to get a variety of atoms either in precomputed matrix (FFT, PCA, etc) or a “dictionary” which depicts from a train set. The premise remains same to get a superposition of atoms such that it would have the most exact and accurate information from the images and aware of which atoms to use, and to have a good representation of image as well. Since are linearly independent from each other, the input and output images relationship is not as linear as other transformations.

VI. Conclusion

CNN architecture as a deep learning method has been proposed to recognize human activity through video dataset. Due to the automatic feature extraction of input data, the proposed CNN architecture achieved the highest accuracy rate and could improve the performance of the classification of human action when compared to the previous studies which have used conventional feature extraction and machine learning methods on the same dataset. Moreover, in this study to achieve accurate results instead of using the commonplace CNN architectures, a specific architecture of CNN has been proposed to use in the dataset. Therefore, this study is the first study on dataset which has applied deep learning models.

VII. Future Work

The Human Image Threshing (HIT) machine uses a mask region-based convolutional neural network (R-CNN) effectively for human body detection, a facial image threshing machine (FIT) for image cropping, resizing, and a deep learning model for activity classification. The HIT machine achieved 98.53% accuracy when the ResNet architecture used as its deep learning model.

Reference

1. TahminaZebin, Patricia J. Scully and Krikor B. Ozanyan, "Human Activity Recognition with Inertial Sensors using a Deep Learning Approach," in proc. 2016 IEEE SENSORS, Orlando, FL, USA. Doi: 10.1109/ICSENS.2016.7808590.
2. Muhammad Mubashir,Ling Shao and LukeSeed, "A survey on fall detection: Principles and approaches," Neurocomputing, vol. 100, p. 144–152, 2013. Doi: 10.1016/j.neucom.2011.09.037.
3. ParisaRashidi and Diane J. Cook, "Keeping the Resident in the Loop: Adapting the Smart Home to the User," IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 39, no. 5, pp. 949-959, Aug. 2009. Doi: 10.1109/TSMCA.2009.2025137.
4. Liming Chen, Jesse Hoey, Chris D. Nugent, Diane J. Cook and Zhiwen Yu, "Sensor-Based Activity Recognition," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 6, pp. 1-19, May 2012. Doi: 10.1109/TSMCC.2012.2198883.
5. Shuiwang Ji, Wei Xu, Ming Yang and Kai Yu, "3D Convolutional Neural Networks for Human Action Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221 - 231, March 2013. Doi: 10.1109/TPAMI.2012.59.
6. Ivan Laptev and Tony Lindeberg, "Space-time interest points," in proc. 2003 Ninth IEEE International Conference on Computer Vision, Nice, France. Doi: 10.1109/ICCV.2003.1238378.