Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
VOLUME:7, ISSUE:2, PAPER NUMBER:53

Human Action Recognition in Videos

Radha K[1], Sarmila K B[2], Vijayakumar S[3], Hariharasudhan D[4]

[1] Faculty and [2,3,4] Students
Dept. of Computer Science Engineering,
Dr. Mahalingam College of Engineering and Technology, Pollachi, India.
radhak@drmcet.ac.in

Abstract: Human action recognition (HAR) plays a critical role across numerous sectors like medical care, recovery support, and support for elders. Utilizing data from mobile sensors such as accelerometers and gyroscopes, researchers employ a variety of Machine Learning or Deep Learning (DL) techniques. DL, especially, streamlines the extraction of intricate features, elevating the effectiveness of HAR systems. The adoption of DL methods has yielded considerable success in HAR across various domains. Recognizing human actions ishighly significant across various sectors, emphasizing its crucial role in enabling intelligent systems to accurately perceive and interpret human behavior. The proposedmodel aims to present a holistic approach for human action recognition in videos utilizing the ResNet-32architecture pre-initialized on the Kinetics dataset.
Keywords— *Human Action Recognition; ResNet-32;Kinetics dataset; Deep Learning*

1. INTRODUCTION:

Human action recognition in videos entails automatically discerning and categorizing human activities depicted in video footage. It encompasses the extraction of features from video frames, including aspects like motion, shape, and appearance. Machine learning algorithms are then employed to classify these features into distinct action types. Its applications span from enhancing video surveillance to analyzing sports activities, empowering systems to comprehend and interpret human actions in real-world contexts. Widely utilized techniques such as support vector machines, deep neural networks, and temporal smoothing contribute to refining accuracy. Preprocessing steps are implemented to ready the raw video data for analysis, while post-processing methods fine-tune classification outcomes. Overall, human action recognition in videos serves as a pivotal tool across various domains, encompassing healthcare monitoring, human- computer interaction, and entertainment.

Recognizing human actions in videos encounters numerous obstacles, such as variations in lighting, background clutter, occlusions, and viewpoint changes, impacting the accuracy of feature extraction and classification. The presence of ambiguities in action boundaries and temporal dynamics further complicates the task. Challenges arise from the scarcity of training data and necessity for extensive annotated datasets to build robust models. Meeting real-time processing demands necessitates the development of algorithms for prompt action recognition. Moreover, effectively handling intricate actions with different poses.

Adapting to diverse environments and user specific actions introduces additional complexities. Additionally, ensuring scalability and generalization across various scenarios remains an ongoing challenge.

In Human Action Recognition (HAR), ResNet operates on frames of the videos via a deep neural network, extracting significant visual features that encompass spatial information vital for action comprehension. Subsequently, recurrent or temporal convolutional networks are employed to model temporal dependencies within videos. These networks scrutinize the sequence of features over time, capturing the evolving dynamics of actions. Finally, a classification layer is incorporated to forecast the action label for the entire video.The main aim is to present a holistic approach for recognizing actions of humans in videos utilizing the ResNet-32 architecture pre-initialized on the Kinetics dataset.

The following sections of this document are organized as follows: Section II offers a summary of the current state-of- the-art in the relevant field. The next section Section III, the proposed approach for recognizing actions in videos, covering aspects such as the dataset pre-processing techniques, algorithms, and evaluation metrics is discussed. Section IV delves into the evaluation results obtained from the conducted experiments. Finally, Section V serves as the conclusion of the paper.

2. RELATED WORK:

Recent researchers have developed various strategies to recognize actions in videos. Below is a synopsis of the latest pertinent works at the forefront of this field.The author introduces a two-stream CNN architecture for action recognition. Demonstrates the importance of bothspatial and temporal information. The architecture is linked to the two streams hypothesis, suggesting that the visual cortex of human consist of two ways: the recognition of objects using ventral stream and the perception of motion by dorsal stream. [1]

Morsheda Akter introduces a method that integrates characteristics from various convolutional stages to create a more extensive feature portrayal. Additionally, it includes an attention mechanism to capture finer features, thereby improving the models accuracy. [2]

Limin proposed a temporal segment network framework for capturing long-range temporal structures. Utilizes 3 Dimensional convolutional networks to process video segments. Achieves competitive performance on action recognition tasks. Through the utilization of the recommended RGB difference for motion models,our method sustains an competitive accuracy on UCF101 (91%) while achieving a processing speed of 340 FPS. [3]

Heng Wang introduces improved dense trajectories for action recognition. Utilizes handcrafted features and trajectory alignment techniques. Achieves robust performance on challenging action benchmarks. [4]

DeepGRU, uses the skeleton of the person, data or the posture of the person that is understood quickly. A model compatible with raw vector data that possesses qualities such as being straightforward to implement, user-friendly, and functions effectively with noisy data. It's also simple to train without demanding high-end hardware, yet manages to attain state-of-the-art outcomes across multiple applications, even with minimal training data. [5]

Introduces 3D CNNs for recognizing actions in humans to directly capture spatiotemporal features. Demonstrates improved performance compared to 2D CNNs on action recognition tasks. Highlights the importance of modeling temporal dynamics in video data. [6]

This model detects and make the action into sections using temporal convolutional network. Demonstrates the effectiveness of 1D convolutions for temporal modeling. Achieves competitive results on action segmentation benchmarks. [7]

Noor Almaadeed, in this model represent data by the method that includes the individuals image extraction in sequence from the scenario. The findings of this study demonstrate that the proposed approach delivers precise recognition of multiple actions of the humans. [10]

This model introduces a hybrid search methodology to train the convolutional neural network classifier weight. This method leverages the effective broad and narrow search capabilities of evolutionary and classical optimization algorithms for accurately predicting activities of humans in uncontrolled video settings. [11]

3. FLOWCHART:

In our proposed system, we introduce a detailed framework for recognizing human actions in videos, utilizing the ResNet-32 architecture pre-trained on the Kinetics dataset. This section outlines the essential components and methodologies utilized in our approach.

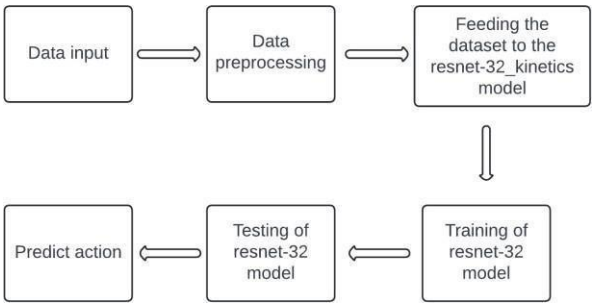

Fig. 3.1. Flowchart diagram for human action recognition

A.DATA PREPROCESSING:

Prior to model training, the video data undergoes preprocessing to standardize the format and facilitate efficient processing. Each video is decomposed into
individual frames, and temporal information is preserved through frame sequencing. We apply standard normalization techniques to ensure uniformity in pixel intensities across frames, thus mitigating the impact of lighting variations and enhancing model generalization.

B.TRANSFERING LEARNING WITH RESNET-32 KINETICS:

Transfer learning serves as a cornerstone in our approach, enabling the utilization of pre-trained deep neural networks to expedite model convergence and enhance performance. The ResNet-32 architecture, pre-trained on the Kinetics dataset, encapsulates rich spatiotemporal features relevant to human action recognition. By leveraging this pre- trained model, we capitalize on its ability to extract discriminative features from videos, thereby circumventing the need for extensive data annotation and model training from scratch.

C.REFINEMENT STRATEGY:

To customize ResNet-32_Kinetics structure to the target action recognition task, we employ a fine- tuning strategy. It involves retraining this model in a task-specific data collection while preserving learned representations from the pre-training phase. During fine-tuning, the model parameters are updated using back propagation with a reduced learning rate to prevent catastrophic forgetting and ensure retention of essential features. This process allows the model to specialize in discriminating between various human actions present in the target dataset.

D.TRAINING PROTOCOL:

During the training phase, the fine-tuned ResNet- 32_Kinetics model is enhanced using a stochastic gradient descent (SGD) algorithm with force. We adopt a batch- wise training approach to accommodate large-scale video datasets efficiently. Furthermore, expanding the dataset by using the techniques such as random cropping, flipping, and temporal jittering are used to augment the training data.

E.MODEL EVALUATION:

Following model training, the performance of the proposed system is analyzed using the performance measures like classification accuracy, precision, recall, and F1-score. We carry out comprehensive tests on both training and test datasets to assess the capability of the model across diverse action categories and video sequences.

F.COMPUTATIONAL EFFICIENCY:

Efficient utilization of computational resources is a paramount concern in deploying deep learning models for real- world applications. We enhance the inference pipeline efficiency by utilizing hardware accelerators like GPUs or TPUs, speeding up model inference while maintaining accuracy. Additionally, techniques like model pruning and quantization may be explored to reduce the model's memory footprint and inference latency.

4. PERFORMANCE EVALUATION AND RESULTS:

The paper emphasized the significance of Data Pre- processing, focusing on the utilization of the ResNet-32 Kinetics dataset. This dataset comprises a diverse array of human actions, with 13,000 videos categorized into 400 action classes, spanning a complete time of 27 hours. Every video ranges from 4 to 11 seconds in length, providing sample variability for model training and testing. Noteworthy categories of action within the dataset include Cricket Bowling, High Jump, Playing Piano, Yoyo, and many more.

Efficient data pre-processing is crucial for enhancing model accuracy and efficiency. Noise within the dataset can hinder model performance, thus necessitating cleaning techniques to improve accuracy. By employing sophisticated pre-processing methods, the accuracy of CNN model range from 80-88% on the ResNet-32 Kinetics dataset.

During training and testing, various combinations of images were used, ranging from 6000 to 12000, to train the ResNet-32 Kinetics model. When the number of images ranges from 20-40% of the frame rate per second then the optimal result is obtained with video-to-image conversion strategies. For instance, in a small frame video, selecting every third frame was found to yield the best results.

Moreover, as the number of times the algorithm process through the dataset increases, both accuracy and loss demonstrated improvements. However, beyond a certain number of times process through the dataset, the increase in accuracy and decrease in loss will be in a stable state.

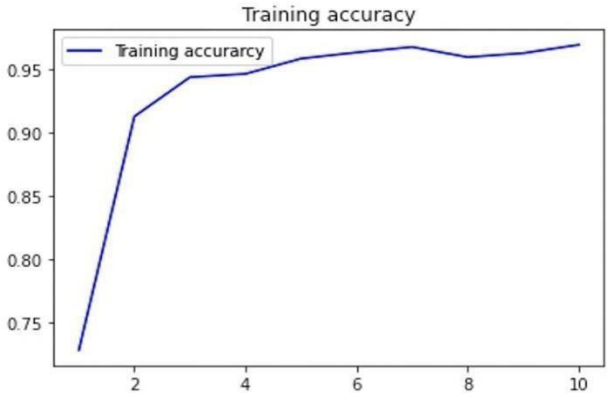

Fig 5.1 Accuracy obtained during training

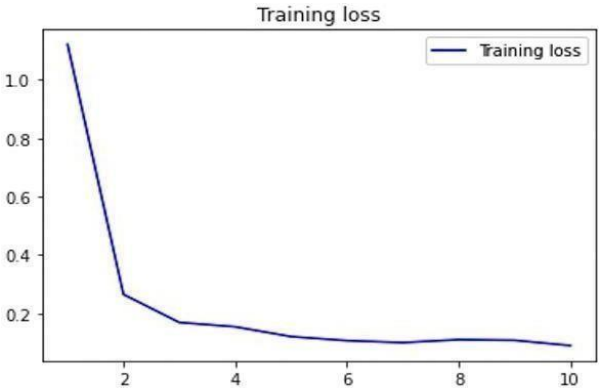

Fig 5.2 Loss obtained during training

After CNN model is trained using different sets of training data comprising of images, evaluation revealed that level of accuracy ranges from 80% to 88%. Subsequently, accuracies were computed for all action classes, indicating that classes with lower video resolution and quality exhibited low level of accuracy. It was analyzed that categories of actions sharing similar activities, especially those with overlapping features, showed slight errors in classification.

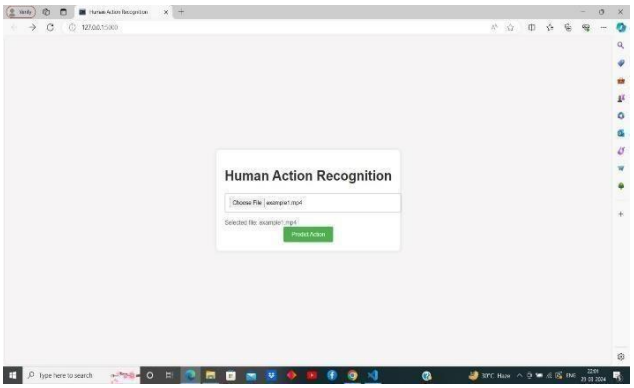

Fig 5.3 Home page

Action recognition for push-ups entails training deep learning models, like ResNet-32_Kinetics, on datasets containing push-up videos. By leveraging convolutional neural networks and transfer learning, the system identifies and classifies push-up motions accurately. This facilitates applications in fitness monitoring, sports analysis, and healthcare.

Fig 5.4 Action recognition

5. CONCLUSION:

The human activity recognition system we've developed utilizes a Convolutional Neural Network (CNN) trained on the Kinetic dataset, enabling precise identification of close to 400 human activities. This system provides the ability to autonomously categorize video datasets, aid in employee training and supervision for task compliance, verify the quality of food service, and oversee patrons in bars/restaurants to ensure satisfactory service. To broaden its applicability, future endeavors could focus on incorporating datasets encompassing more than 400 activities. Additionally, we've observed that increasing the quantity of samples per activity in the data collection, boosts the system performance.

6. REFERENCES:

[1] Simonyan, K., and Zisserman, A. (2014). "Utilizing two- stream convolutional networks for action recognition in videos." Published in Advances in Neural Information Processing Systems, Volume 27.

[2] Akter M, Ansary S, Khan MA, Kim D. "Human Activity Recognition through Deep Learning Feature Combination with Attention Mechanisms." Published in Sensors (Basel), June 19, 2023; Volume 23, Issue 12, Pages 5715. DOI: 10.3390/s23125715. PMID: 37420881; PMCID: PMC10301803.

[3] Wang, Limin, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. "Action Recognition in Videos Using Temporal Segment Networks." Published in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017; Volume 41, Pages 2740-2755.

[4] Wang, H., and C. Schmid. "Enhanced Trajectories for Action Recognition." Presented at the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia. Published in Proceedings, Pages 3551- 3558. DOI: 10.1109/ICCV.2013.441.

[5] Maghoumi, M., and LaViola, J. J. (2019). "DeepGRU: A Utility for Deep Gesture Recognition." Presented at the 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7–9, 2019.

[6] Ji, Shuiwang, Xu, Wei, Yang, Ming, and Yu, Kai. (2010). "Utilizing 3D Convolutional Neural Networks for Human Action Recognition." Published in IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume 35, Pages 495-502. DOI: 10.1109/TPAMI.2012.59.

[7] Lea, C., Flynn, M. D., Vidal, R., Reiter, A., and Hager,G.D. (2017). "Temporal convolutional networks for action segmentation and detection." Presented at the IEEE Conference on Computer Vision and Pattern Recognition (pp. 156-165).

[8] Wang, Wenguan, Zhou, Tianfei, Porikli, Fatih, Crandall, David, and Gool, Luc. (2021). "An Overview of Deep Learning Methods for Video Segmentation."

[9] Kong, Y., and Fu, Y. (2022). "A Comprehensive Review of Human Action Recognition and Prediction."Published in the International Journal of Computer Vision, 130(5), 1366-1401.

[10] Almaadeed, N., Elharrouss, O., Al-Maadeed, S., Bouridane, A., & Beghdadi, A. (2019). A novel approach for robust multi human action recognition and summarization based on 3D convolutional neural networks. arXiv preprint arXiv:1907.11272.

[11] Paul, Earnest & Chalavadi, Krishna Mohan. (2016). Human action recognition using genetic algorithms and convolutional neural networks. Pattern Recognition. 59. 199-212. 10.1016/j.patcog.2016.01.012.

[12] A.Iosifidis, A. Tefas and I. Pitas, "Multi-view action recognition based on action volumes fuzzy distances and cluster discriminant analysis", Signal Processing, vol. 93, no. 6, pp. 1445-1457, Jun. 2013.

[13] D. Weinland, R. Ronfard and E. Boyer, "A survey of vision-based methods for action representation segmentation and recognition", Comput. Vis. Image Underst, vol. 115, no. 2, pp. 224-241, Feb. 2011.

[14] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning", IEEE Trans. Pattern Anal. Mach. Intell, vol. 32, no. 2, pp. 288-303, Feb. 2010.

[15] A.Oikonomopoulos and M. Pantie, "Human Activity Recognition Using Hierarchically-Mined Feature Constellations", pp. 150-159, 2013.

[16] Javan Roshtkhari and M. D. Levine, "Human activity recognition in videos using a single example", Image Vis. Comput., vol. 31, no. 11, pp. 864-876, Nov. 2013.

[17] J. Yang, J. Lee and J. Choi, "Activity Recognition Based on RFID Object Usage for Smart Mobile Devices", J. Comput. Sci. Technol., vol. 26, no. 2, pp. 239-246, Mar. 2011.

[18] L. Chen, H. Wei and J. Ferryman, "A survey of human motion analysis using depth imagery", Pattern Recognit. Lett., vol. 34, no. 15, pp. 1995-2006, Nov. 2013.

[19] W. Ong, L. Palafox and T. Koseki, "Investigation of Feature Extraction for Unsupervised Learning in Human Activity Detection", Bull. Networking Comput. Syst. Softw, vol. 2, no. 1, pp. 30-35, 2013.

[20] O. D. Lara and M. A. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors", IEEE Commun. Surv. Tutorials, vol. 15, no. 3, pp. 1192- 1209, Jan. 2013.

[21] A. A. Chaaraoui, J. R. Padilla-López, P. Climent-Pérez and F. Flórez-Revuelta, "Evolutionary joint selection to improve human action recognition with RGB-D devices", Expert Syst. Appl., vol. 41, no. 3, pp. 786- 794, Feb. 2014.

[22] N. Noorit and N. Suvonvorn, "Human Activity Recognition from Basic Actions Using Finite State Machine", Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng- 2013), vol. 285, pp. 379-386, 2014.

[23] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos", 2011 International Conference on Computer Vision, pp. 1036-1043, 2011 .