Convolutional Neural Networks for Human Action Recognition

Ahmed Ali Ali Mahmoud
Artifical intelligence
Cairo university
Cairo , Egypt
aly869556@gmail.com

Ibrahim Hisham Ahmed
Artifical intelligence
Cairo university
Cairo , Egypt
ibrahimhishamzz23@gmail.com

***Abstract*— This proposal outlines a research project aimed at exploring the application of Convolutional Neural Networks (CNNs) for Human Action Recognition (HAR). The proposed method employs the Mobile Net CNN architecture and leverages pretraining on the ImageNet dataset. The study utilizes a carefully curated dataset consisting of 15 distinct human actions captured in various scenarios, contributing to a total of 12,600 samples. The dataset is designed to cover a broad spectrum of activities, including sitting, using a laptop, hugging, sleeping, drinking, clapping, dancing, cycling, calling, laughing, eating, fighting, listening to music, running, and texting.

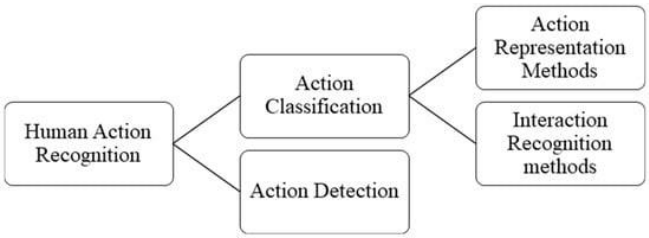
I. INTRODUCTION

Human action recognition has a wide range of applications, video storage and retrieval [1,2] , identity recognition [3], intelligent human–machine interfaces [4,5] , video surveillance and environmental home monitoring [6,7].

Human Action recognition covers many research topics in computer vision including human detection in video (radars), human tracking and understanding time series data, it's also challenging in field computer vision and ML.

The key to good human action recognition is robust human action modeling and feature representation. Feature representation and selection is a classic problem in computer vision and machine learning [8]. Unlike feature representation in an image space, the feature representation of human action in video not only describes the appearance of the human(s) in the image space, but must also extract changes in appearance and pose. The problem of feature representation is extended from two-dimensional space to three-dimensional space-time.

Human action recognition is centered on the detection and tracking of human movements, coupled with the classification or categorization of actions depicted in image datasets. This task poses a formidable challenge, demanding the development of robust and efficient techniques aimed at minimizing error rates. The intricacy lies in the dynamic nature of human actions within single images and the diverse array of scenarios captured. Consequently, researchers and practitioners in this field continuously strive to advance methodologies that not only enhance accuracy but also contribute to the broader domains of computer vision and artificial intelligence. The pursuit of precision in human action recognition is fundamental to unlocking its full potential across various applications, from surveillance to human-computer interaction within static image datasets.



II. RELATED WORK

Handcrafted feature-based action representations have dominated the computer vision field alongside action recognition. Before the emergence of deep learning approaches, most of the handcrafted feature methods used in action recognition employed a fixed procedure, including feature extraction, feature representations, and action classification.
In this approach, low-level action features, such as HOF, HOG, and STIP, are extracted from those sparse spatial temporal features, and projected to video-level representations through coding methods, such as the bag-of words (BOW) model, and sparse coding (SC). However, some later works prove that improved performance can be achieved by adopting densely sampled spatial-temporal features.

Wang et al. proposed an action recognition method with improved dense trajectories (IDT) that integrates SURF and optical flows. Their method improved motion based descriptors significantly. Motivated by the success of IDT, researchers are working intensely towards developing IDT for video-feature learning. In, the performance of IDT is enhanced by building a multi-layer stacked Fisher Vector (FV) method. The authors of proposed a method to sub-sample and generate vocabularies for DT features. In, the performance of base action classifiers is improved by adopting three methods: data augmentation, Subsequence-Score Distribution, and Least-Squares SVMs. The authors of proposed an approach called Multi-skip Feature Stacking

(MIFS), and achieved state-of-the-art results on several datasets. In 2016, Lan et al. proposed leverage effective techniques from both data-driven and data-independent approaches, to improve action recognition systems. However, although the extraction of more spatial-temporal features from the video improves performance, it also incurs higher computational costs.

.

A. Experimental Dataset

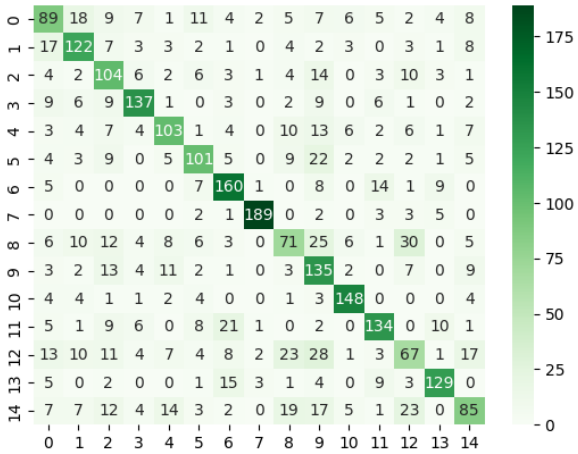
In our experiment, we test our method on the Human Action Recognition (HAR) []as shown in Fig. 2. The Images dataset containing 15 types of human actions (Sitting, using laptop, hugging, sleeping, drinking, clapping, dancing, cycling, calling , laughing , eating , fighting , listing to music , running and texting) Each class of them contain 840 Sample with different situations and persons . Currently the Dataset contains 12600 sample. All Samples were taken over homogeneous backgrounds with a static camera . Each image has the shape of 160 * 160 pixels with 3 channels.

For data preparation, first, we retrieve each image from its path, input it into an array, and assign its label to another array. Next, we convert the label array into numbers corresponding to each category, ranging from 0 to 14. Then, we split the data with a ratio of 78% for the training set and 22% for the test set.

B. Experimental Results

In the pre-training CNN-based HAR system, we train a Mobile Net ىv2 and employ it to extract features. We then utilize the K-best optimizer in the feature selection stage and proceed to perform classification using a Support Vector Machine, resulting in an accuracy of 64% the following

	precision	recall	f1-score
0	0.51	0.50	0.51
1	0.65	0.69	0.67
2	0.51	0.64	0.57
3	0.76	0.74	0.75
4	0.66	0.60	0.63
5	0.64	0.59	0.62
6	0.69	0.78	0.73
7	0.95	0.92	0.94
8	0.47	0.38	0.42
9	0.46	0.70	0.56
10	0.83	0.86	0.84
11	0.73	0.68	0.70
12	0.42	0.34	0.38
13	0.79	0.75	0.77
14	0.56	0.43	0.48
accuracy			0.64
macro avg	0.64	0.64	0.64
weighted avg	0.64	0.64	0.64



figures showing performance measures and confusion matrix.

Proposed model

1. Implementation of Mobile Net CNN:

The foundation of the proposed model lies in the utilization of the Mobile Net CNN architecture for efficient feature extraction in the domain of Human Action Recognition (HAR). The MobileNetV2 model is configured with weights pretrained on the ImageNet dataset, ensuring the extraction of meaningful and discriminative features from input images with dimensions (160, 160, 3).

2. ImageNet Pretraining:

To enhance the model's ability to discern high-level features relevant to HAR, ImageNet pretraining is employed. By initializing the MobileNetV2 model with weights learned from ImageNet, the proposed model gains a rich set of hierarchical features, facilitating a more nuanced understanding of human actions.

3. Performance Evaluation:

The effectiveness of the proposed method is rigorously assessed through a comprehensive performance evaluation on the HAR dataset. Metrics such as accuracy, precision, recall, and F1 score are employed to provide a holistic view of the model's classification capabilities. This evaluation is crucial for gauging the model's robustness and suitability for real-world applications.

4. Comparative Analysis:

A critical aspect of the proposed research involves a comparative analysis with existing HAR methods. By benchmarking the proposed Mobile Net CNN architecture against established approaches, the model's effectiveness and efficiency are thoroughly examined. This comparative analysis aims to identify the unique contributions and

potential advancements introduced by the proposed model within the field of HAR.

5. Identification of Challenges:

Acknowledging the diverse challenges and limitations in the application of CNNs to HAR is integral to the proposed model. An in-depth exploration is conducted to identify and document challenges, considering the unique characteristics of the dataset. This analysis contributes valuable insights that guide the refinement and optimization of the proposed CNN architecture for improved human action recognition.

6-Feature Selection:

The subsequent code snippet selector = SelectKBest(score_func=mutual_info_classif, k=k_best) utilizes the SelectKBest method from the scikit-learn library. This method employs a specified scoring function (mutual_info_classif in this case) to evaluate the importance of each feature in the dataset. The k parameter determines the number of top features to be retained based on their scores. This process aids in reducing dimensionality and focusing on the most informative features, contributing to improved model efficiency and interpretability.

7-Support Vector Machine (SVM):

The following lines svm_model = SVC(kernel='poly') initialize a Support Vector Machine (SVM) model with a linear kernel. SVM is a supervised learning algorithm utilized for classification and regression tasks. In the context of this model, the poly kernel is chosen, but users have the flexibility to experiment with different kernel options based on the nature of their specific problem.

8-Model Training:

The final line svm_model.fit(X_train_selected, y_train) signifies the training phase of the SVM model. Here, X_train_selected represents the subset of features selected through the previous feature selection process, and y_train corresponds to the corresponding labels. The SVM model learns to classify and predict based on the selected features, leveraging the linear kernel in this instance. This step is crucial for the model to generalize patterns from the training data and make accurate predictions on new, unseen data.

conclusion

this research proposal delineates a comprehensive exploration into the realm of Human Action Recognition (HAR) using Convolutional Neural Networks (CNNs). By adopting the MobileNet CNN architecture and incorporating pretraining from the ImageNet dataset, the study aims to enhance the model's capacity for discerning intricate features crucial for accurate action classification. The dataset, meticulously curated to encompass 15 diverse human actions across a myriad of scenarios, serves as a robust foundation for training and evaluating the proposed methodology.

The inclusion of common daily activities, such as sitting, using a laptop, hugging, sleeping, drinking, clapping, dancing, cycling, calling, laughing, eating, fighting, listening to music, running, and texting, within the dataset reflects a commitment to capturing a broad spectrum of human actions. With a substantial dataset comprising 12,600 samples, each exhibiting unique situations and individuals, the study aspires to push the boundaries of HAR research.

The decision to employ a static camera against homogeneous backgrounds ensures consistency in the dataset, allowing for a focused examination of the CNN model's ability to generalize across diverse scenarios. The resolution of 160x160 pixels and the RGB representation with 3 channels contribute to the dataset's richness, providing a nuanced view of human actions in a computationally efficient manner.

As the proposed research unfolds, it is anticipated that the outcomes will not only shed light on the efficacy of the MobileNet CNN architecture in conjunction with ImageNet pretraining for HAR but also offer valuable insights into the broader applicability of CNNs in recognizing a wide array of human actions. The potential implications of this research span diverse domains, including surveillance systems, human-computer interaction, and beyond, with the aim of advancing the state-of-the-art in HAR technology.

REFERENCES

[1] Van Gemert, J.C.; Jain, M.; Gati, E.; Snoek, C.G. APT: Action localization proposals from dense trajectories. In Proceedings of the British Machine Vision Conference 2015: BMVC 2015, Swansea, UK, 7–10 September 2015; p. 4. [Google Scholar].

[2] Zhu, H.; Vial, R.; Lu, S. Tornado: A spatio-temporal convolutional regression network for video action proposal. In Proceedings of the CVPR, Venice, Italy, 22–29 October 2017; pp. 5813–5821. [Google Scholar].

[3] Paul, S.N.; Singh, Y.J. Survey on Video Analysis of Human Walking Motion. Int. J. Signal Process. Image Process. Pattern Recognit. 2014, 7, 99–122. [Google Scholar] [CrossRef].

[4] Papadopoulos, G.T.; Axenopoulos, A.; Daras, P. Real-time skeleton-tracking-based human action recognition using kinect data. In Proceedings of the International Conference on Multimedia Modeling, Dublin, Ireland, 6–10 January 2014; pp. 473–483. [Google Scholar].

[5] Presti, L.L.; Cascia, M.L. 3D Skeleton-based Human Action Classification: A Survey. Pattern Recognit. 2016, 53, 130–147. [Google Scholar] [CrossRef].

[6] Aggarwal, J.K.; Ryoo, M.S. Human activity analysis: A review. ACM Comput. Surv. 2011, 43. [Google Scholar] [CrossRef].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[8] Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 2013, 35, 1798–1828. [Google Scholar] [CrossRef] [PubMed][Green Version].

[9] Ouyang X, Xu S, Zhang C, Zhou P, Yang Y, Liu G, Li X (2019) A 3D-CNN and LSTM based multi-task learning architecture for action recognition. IEEE Access 7:40757–40770

[10] Human Action Recognition based on Convolutional Neural Networks with a Convolutional Auto-Encoder Chi Geng1, a, JianXin Song1, b

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classi_cation with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097{1105, 2012.

[12] A. Tomas and K. Biswas. Human activity recognition using combined deep architectures. In Signal and Image Processing (ICSIP), 2017 IEEE 2nd International Conference on, pages 41{45. IEEE, 2017.

[13] Zou C, Kou KI,Wang Y (2016) Quaternion collaborative and sparse representation with application to color face recognition. IEEE Trans Image Process 25(7):3287–3302

[14] Ji, S., Wei, X., Yang, M., Kai, Y.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)