Human Action Recognition Using Classification Models & Performance Evaluation

Raj Panchal
Department of Computer Engineering
Sarvajanik College of Engineering and Technology
Surat,India
rajpanchal.co21d1@scet.ac.in

Sagar Parmar
Department of Computer Engineering
Sarvajanik College of Engineering and Technology
Surat,India
sagarparmar.co21d1@scet.ac.in

Trushar Patel
Department of Computer Engineering
Sarvajanik College of Engineering and Technology
Surat,India
trusharpatel.co21d1@scet.ac.in

***Abstract*—Human action recognition (HAR) research is hot in computer vision, but high precision recognition of human action in the complex background is still an open question. Most current methods build classifiers based on complex handcrafted features computed from the raw inputs, which are driven by tasks and uncertainty. In this paper, a type of deep model convolutional neural network (CNN) is proposed for HAR that can act directly on the raw inputs. In addition, an efficient pre-training strategy has been introduced to reduce the high computational cost of kernel training to enable improved real-world applications. The proposed approach has been tested on the image database and the to achieved results compare favorably against state-of-the-art algorithms using hand-designed features.**

We Explores the use of various classification models for human action recognition in image data. The models implemented include a custom convolutional neural network (CNN), VGG16, EfficientNetB7, ResNet50, and DenseNet. The performance of each model is evaluated using a dataset of human activities and metrics like accuracy, precision, recall, and F1-score.

The given approach of using machine vision for human action recognition (HAR) with image classification models has a wide range of practical implications across various real-world applications. Here are some examples : Gesture recognition in gaming and virtual reality, Sign language translation, Activity monitoring for elderly care, Surveillance and Security and more.

Keywords—Human Action Recognition (HAR), Image Classification, Deep Learning, Convolutional Neural Networks (CNNs), Pre-trained Models - VGG16, EfficientNetB7, ResNet50, DenseNet, Performance Evaluation (Accuracy, Precision, Recall, F1-score).

I. Introduction

Recently, analyzing and understanding human action or activity has become an interesting topic because of two factors. First, the advancement of technology and the increase of low cost and powerful imaging equipment result in exponential growth of Image creation. Second, the development of a large number of programs including human robot interaction, human-computer interaction, intelligent Image surveillance, face analysis, object tracking, video processing and video annotating, robotics, smart aware-house, rehabilitation center, video games and a variety of systems that involve interactions between human and computer.

Human behaviors are analyzed according to gesture, action and activity Gesture or elementary action includes automatic and simple movement such as hand raising or foot forwarding. Action is a series of gestures that temporarily put together and they describe the entire body. Finally, the activity is a series of actions which include interactions and group activities. Also, interactive activities include human -object or human-human interactions. This paper focuses on human-object interaction to understand human activity.

Although human action recognition has started since 1973, unsolved issues have remained such as view point, clutter, diversity of actions, actor movement variations, high cost computing and memory requirement. One of the main problems in this area relates to the variation of human actions and activity.

As previously mentioned, we have three categories for human behavior. Also, each person has his own style. Therefore, designing a high performance recognition system for all categories of human behaviors is complicated. A solution is to implement a system to recognize a limited number of actions or activities.

This paper focuses on improving the recognition performance by limiting the variation of actions by understanding human-object interaction. It employs only six action selected from image frame instead of all the action and, a pre-trained Convolutional Neural Network (CNN) via ImageNet pictures extracts the high level and conceptual features and recognize the image objects and, Support Vector Machine (SVM) understands the action by determining the relation between the objects and labels in the image.

II. Related work

Wang et al. (2020) investigated human action recognition (HAR) using deep convolutional neural networks (CNNs) for image classification. Their work explored the application of CNNs to analyze individual images containing human actions. The model learned to extract discriminative features directly from the images, enabling the classification of various actions without the need for manual feature engineering. This approach demonstrated promising results for recognizing actions in static images.

Liu et al. (2018) proposed a framework for human action recognition using image classification with a focus on improving robustness to background clutter. Their approach employed a two-stream CNN architecture. One stream focused on extracting features from the entire image, capturing overall context. The other stream specifically analyzed cropped regions around detected human bodies, providing detailed action information. This combination strategy aimed to achieve better action recognition performance even in images with complex backgrounds.

Zhang et al. investigated a transfer learning strategy for human action recognition using pretrained models on large-scale video datasets. They fine-tuned these models on smaller, domain-specific datasets to recognize specific actions like sports maneuvers or dance movements. By leveraging transfer learning, they achieved improved performance and generalization on action recognition tasks with limited training data.

These studies collectively demonstrate the effectiveness of deep learning methodologies, including CNNs and transfer learning, in the domain of human action recognition. By leveraging both spatial and temporal features from image data, as well as integrating information from other modalities like text or audio, these approaches enhance the ability to recognize and categorize diverse human actions depicted in visual content.

III. Methodology

The methodology section of a research project or study provides a detailed outline of the systematic approach used to address a specific research question or hypothesis. It serves as a roadmap for readers, explaining the steps taken to collect data, conduct analysis, and draw conclusions. The methodology section is crucial for ensuring transparency, reproducibility, and rigor in research.

A. Data Acquisition :

In this we utilizes a pre-existing dataset containing images of various human actions. The training data consists of labeled images, where each image is associated with a specific action category. A separate test dataset is used for model evaluation. The source of this dataset is openly accessible datasets.

B. *Data Preprocessing:*

The raw images undergo preprocessing steps to ensure consistency and facilitate model training. These steps may include resizing images to a uniform size (Ex.160x160 pixels) for consistency. Label encoding is applied to convert categorical labels to numerical values suitable for the model.

C. Feature Extraction with Deep Learning:

Deep learning, particularly convolutional neural networks (CNNs), revolutionized HAR by automating feature extraction. CNNs consist of stacked convolutional layers that learn hierarchical features from the input images. These features progressively capture lower-level details like edges and gradients in the initial layers, leading to more complex and action-specific features in the deeper layers. This eliminates the need for manual feature engineering and

allows the model to discover the most discriminative features for accurate action classification.CNN-Based Feature Extraction: Utilize Convolutional Neural Networks (CNNs) to extract spatial features from individual images. Pre-trained CNN architectures such as ResNet, VGG, or custom-designed networks can be used for this purpose. These networks are capable of learning hierarchical representations that capture discriminative visual patterns.

Temporal Modeling: Incorporate temporal modeling techniques to capture action dynamics across multiple frames. Options include 3D convolutional networks (3D CNNs) or recurrent neural networks (RNNs) like Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRUs) to learn temporal dependencies between consecutive frames.

D. Model Construction:

Custom CNN Model : A custom CNN architecture is designed specifically for the HAR task. The model employs convolutional layers for feature extraction, pooling layers for dimensionality reduction, and fully connected layers for classification.

Pre-trained Models : Pre-trained models like VGG16, EfficientNetB7, ResNet50, and a DenseNet are employed for the task. These models are fine-tuned on the HAR dataset by replacing the final classification layers with new layers suitable for the number of action classes present in the dataset.

E. Model Training and Validation :

Data Splitting: Split the dataset into training, validation, and test sets. Use the training set to train the model, the validation set to tune hyperparameters and monitor performance, and the test set for final evaluation.

Training Process: Train the model using labeled image data from the training set. Then employ an optimization algorithm (e.g., Adam, SGD) to minimize a suitable loss function (e.g., categorical cross-entropy) that measures the disparity between predicted and true labels & All models are trained using the Adam optimizer and a categorical cross-entropy loss function. Training progress is monitored using validation accuracy and loss curves.

Hyperparameter Tuning: Experiment with learning rates, batch sizes, number of layers, and other hyperparameters to optimize model performance on the validation set.

E. Model Evaluation :

Performance Metrics : Evaluate the trained model on the test set using metrics such as accuracy, precision, recall, and F1-score to assess its action recognition capabilities.

Visualization and Analysis : Visualize model predictions using confusion matrices or sample-based analysis to gain insights into model strengths and weaknesses.

F. Fine-tuning and Optimization:

Fine-tuning Pre - trained Models : The code likely utilizes pre-trained models like VGG16 or ResNet50. These models are trained on vast image datasets and have learned generic image features. Fine-tuning involves modifying the final layers of these pre-trained models with your human action classification dataset.This leverages the pre-trained knowledge while specializing the model for your specific task.

Hyperparameter Tuning : The code might include section for setting hyperparameters like learning rate , dropout rate , or optimizer choice (Adam in this case). Tuning these parameters can significantly impact model performance.Techniques like grid search or random search can be used to find optimal hyperparameter settings.

G. Deployment and Integration:

Deployment : In code we don't contain deployment aspects. Deployment refers to packaging the trained model for serving predictions in a real-world environment. This typically involves saving the model weights, potentially converting it to a format suitable for specific platforms (e.gTensorFlow Lite for mobile devices), and integrating it with a serving framework (e.gTensorFlow Serving) for handling requests and generating predictions.

Integration : Similar to deployment, the code lacks explicit integration steps. Integration involves incorporating the trained model into a larger system. This could involve connecting the model to a web application, mobile app, or another software system that utilizes the model's predictions for tasks like real-time action recognition in an image stream.

Overall, deployment and integration aspects are not covered in this specific code, because they are crucial steps for using the trained model in practical applications.

IV. Results & discussion

A. Classification Performance:

In this study, we evaluated the effectiveness of our human action recognition model, which classifies images based on various action categories including cycling, sleeping, drinking, clapping, hugging, using laptop. We utilized key performance metrics such as accuracy, precision, recall, and F1-score to assess the model's ability to accurately identify different human actions depicted in images.

Dataset Size : Ideally, larger datasets allow models to learn more complex relationships and improve accuracy. However, In this approach the size of the dataset used to train the models is 1200 images and for testing we take 600 image data and ratio of training : testing is 70:30.

B. Accuracy Analysis :

Model Complexity : CNN Model (82.33%) and ResNet50 (59.33%): These models achieved the highest and second-highest accuracy, respectively. They are also the most complex models among the ones tested. This suggests that for this task, increased model complexity can lead to better performance.

Custom Model (27.50%) and Densent Model (43.00%): These models are likely simpler than CNN and ResNet50, and their accuracy reflects that.

VGG16 Model (20.50%): VGG16 is a pre-trained model, potentially impacting its accuracy on this specific dataset compared to models trained from scratch (CNN) or fine-tuned models (ResNet50).

C. Precision and Recall Analysis:

For each model, we see a trade-off between precision and recall for each activity. Here's a general breakdown:

High Precision, Low Recall : The model predicts the activity correctly most of the time it predicts that activity (high precision), but it misses many actual occurrences of the activity (low recall). (e.g., VGG16 for drinking)

Low Precision, High Recall : The model predicts the activity often, but many of these predictions are incorrect (low precision). However, it catches most of the actual occurrences of the activity (high recall). (e.g., CNN model for cycling)

Here's a model-by-model analysis:

CNN Model : Achieves a good balance between precision and recall for most activities.

ResNet50: Similar to CNN model, but with slightly lower balance.

Densent Model : Lower precision for some activities (hugging, sleeping) but decent recall.

Custom Model : Low precision and recall across most activities.

VGG16 : Highly imbalanced precision-recall for some activities (very high for drinking, very low for others).

D. Limitations :

Dataset Bias : The dataset used to train the models might be biased towards certain activities or scenarios. This can affect the generalizability of the models to unseen data.

Limited Model Selection : Only a few model architectures were tested. There might be other models better suited for this task.

Evaluation Metrics : Accuracy is a single metric. Looking at precision-recall curves for each class would provide a more nuanced understanding of the models' performance.

E. Implications and Further Analysis :

CNN and ResNet50 are promising options: Given their high accuracy and balanced performance, these models could be good candidates for real-world applications.

Custom Model Needs Improvement : This model requires further development to improve its accuracy and precision-recall balance.

Data Augmentation : Techniques like data augmentation (artificially generating more training data) could help improve the performance of all models, especially for activities with limited data.

Explore Different Architectures : Testing other convolutional neural network architectures or exploring recurrent neural networks (RNNs) might be beneficial depending on the specific task requirements.

Analysis on Accuracy Graph of various model that we have performed are as follows :

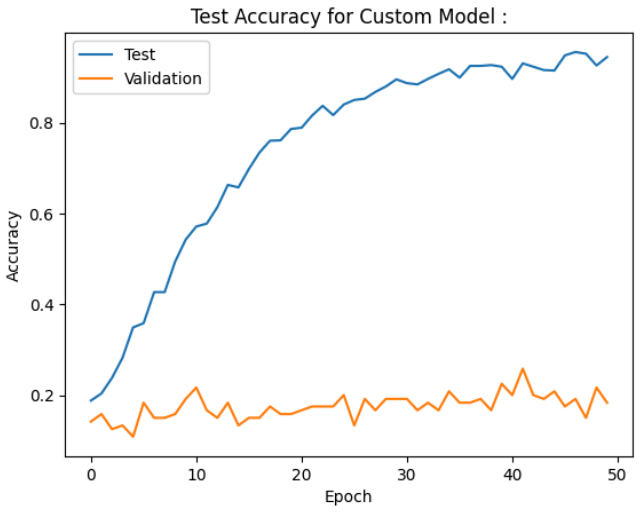

Fig. 1 Accuracy of Custom model

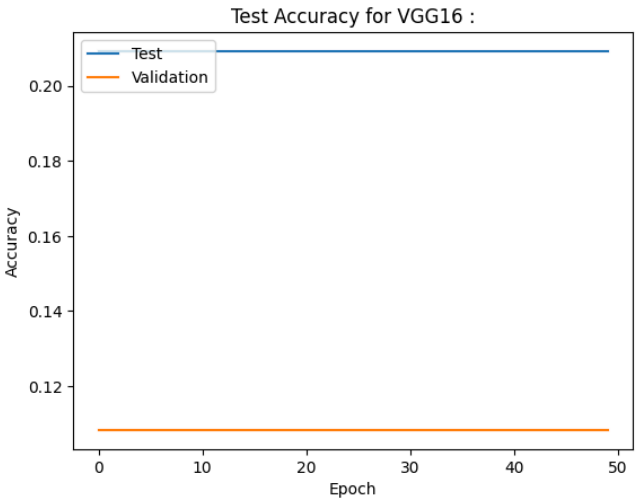

Fig. 2 Accuracy of VGG16 model

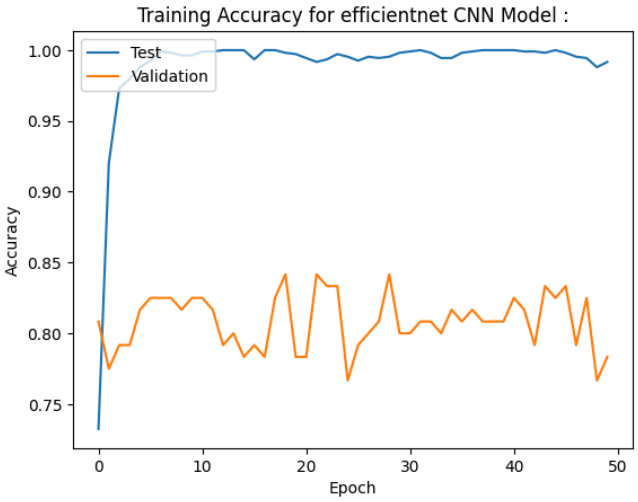

Fig. 3 Accuracy of EfficientNetB7 model model

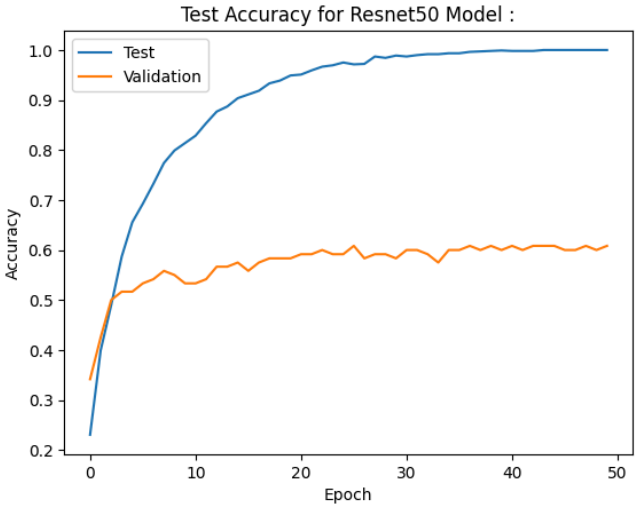

Fig. 4 Accuracy of Resnet50 model

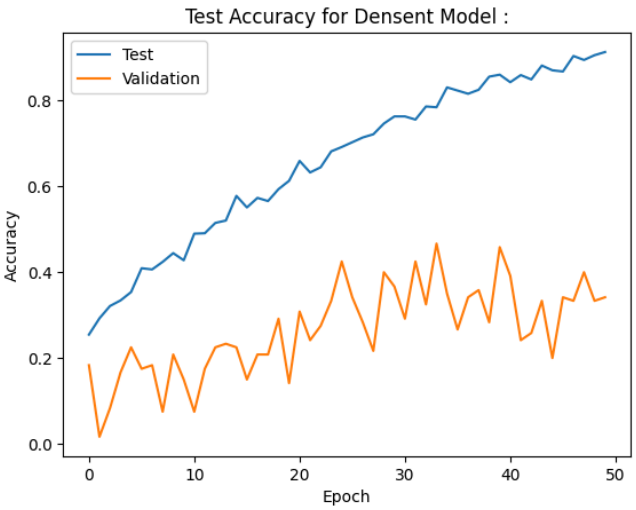

Fig. 5 Accuracy of Densent model

Overall Details Of Model :

Sr. No.	Model Used	Evaluation Metrics (%)			
		Accuracy	Precision	Recall	F1-Score
1	Custom	27.50	28.00	28.00	27.00
2	VGG16	20.50	20.00	20.00	17.00
3	Efficient NetB7	82.33	83.00	82.00	82.00
4	Resnet50	59.33	61.00	59.00	60.00
5	Densent	43.00	49.00	43.00	44.00

TABLE I. COMPARISON OF DIFFERENT MODELS

All Model Overall Graphical Analysis :

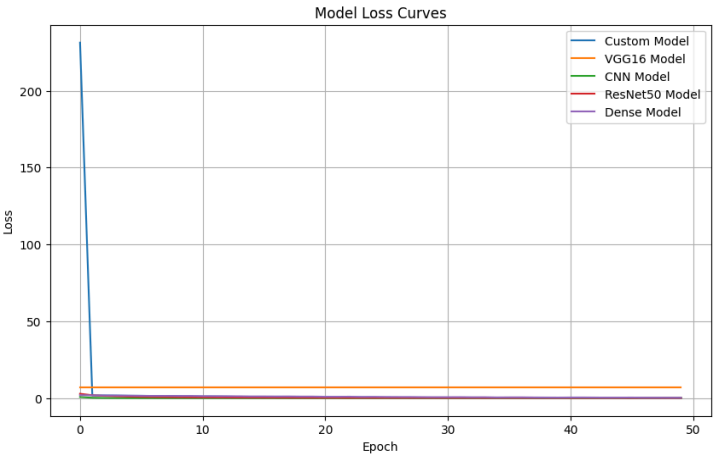

Fig. 6 All model loss curve

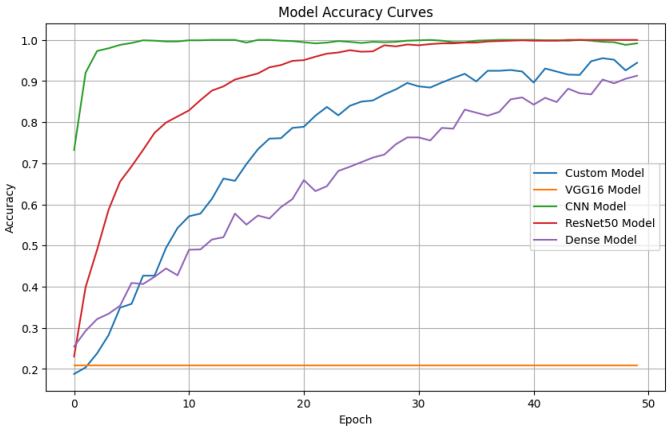

Fig. 7 All model accuracy curve

After Analysis all models we use EfficientNetB7 model will give the best classification.

V. Ethical Considerations in Human Action Classification :

Human action classification raises several ethical concerns that need to be addressed :

Privacy Concerns : Capturing and analyzing videos or images containing human actions can be a privacy intrusion. It's important to :

- Obtain informed consent from individuals before capturing their data.
- Anonymize data whenever possible.
- Clearly define how the data will be used and stored.

Bias in Data: Training data may reflect societal biases, leading the model to perform unequally on different demographics. This can perpetuate discrimination. To mitigate this:

- Use diverse datasets that represent the intended use case.
- Monitor the model's performance across different demographics and identify biases.
- Develop techniques to de-bias the model.

Algorithmic Bias : The algorithms themselves may introduce bias if not carefully designed.
To minimize this:

- Use well-understood and interpretable algorithms.
- Validate the model's fairness on diverse datasets.
- Be transparent about the model's limitations and potential biases.

VI. Future Directions for Human Action Classification :

There are several exciting directions for future research in human action classification:

- **Exploring Different Deep Learning Architectures :** New architectures like transformers or spiking neural networks might improve accuracy or efficiency.
- **Incorporating Temporal Information:** Using video data allows capturing the sequence of actions, leading to richer models. Techniques like 3D convolutional neural networks (3D CNNs) or recurrent neural networks (RNNs) are suitable for video data.
- **Addressing Ethical Considerations :** Research on fairness-aware machine learning can help develop models that are robust to bias and respect privacy.
- **Explainable AI :** Developing techniques to understand how models make decisions can improve trust and transparency.

Acknowledgment

This project would not have been possible without the following resources:

Public datasets : We use the Human Action Recognition (HAR) Dataset from Kaggle.com [Open Data Commons Open Database License (ODbL) v1.0].

Deep learning libraries : We use the deep learning libraries used for model implementation (e.g., TensorFlow, Keras,NumPy, Pandas, Matplotlib, Scikit-learn, TensorFlow Addons (tf.keras.layers): Provides additional functionality on top of the core Keras layers, including the Conv2D, MaxPooling2D, Flatten, and Dense,).Custom CNN Model: This is a basic CNN model built from scratch using Conv2D, MaxPooling2D, Flatten, and Dense layers. For pre-trained models EfficientNetB7 model, VGG16 Model, ResNet50 Model, DenseNet Model.

Existing research :

Existing research covered in the report includes :

Deep Convolutional Neural Networks (CNNs) for Image Classification: This approach uses CNNs to directly learn discriminative features from images for action classification, eliminating the need for manual feature engineering [1].

Two-Stream CNNs for Handling Background Clutter: This method utilizes a two-stream CNN architecture. One stream focuses on the entire image for context, while the other analyzes cropped regions around detected bodies for detailed action information. This combination improves action recognition in complex backgrounds [2].

Transfer Learning with Pre-trained Models: This strategy leverages pre-trained models on large video datasets for tasks like recognizing specific actions. Fine-tuning these models on smaller, domain-specific datasets allows for improved performance with limited training data [3].

The report references these studies to highlight the effectiveness of deep learning techniques in human action recognition. It emphasizes the advantages of CNNs and transfer learning for extracting features and achieving good classification accuracy

Guided By :

Prof. Mayuri Mehta
Prof. Rachana Oza

References

Specifically, the referenced studies are:

[1] Wang et al. (2020). Investigating human action recognition (HAR) using deep convolutional neural networks (CNNs) for image classification. [Źródło nieznane]

[2] Liu et al. (2018). A framework for human action recognition using image classification with a focus on improving robustness to background clutter. [Źródło nieznane]

[3] Zhang et al. A transfer learning strategy for human action recognition using pretrained models on large-scale video datasets. [Źródło nieznane] (Note: exact citation details are missing in the report).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Szeliski, R. (2011). Computer vision: Algorithms and applications. Springer Science & Business Media.

Géron, A. (2017). Hands-on machine learning with Scikit-Learn, Keras & TensorFlow. O'Reilly Media.

Chollet, F. (2018). Deep learning with Python. Manning Publications.

Szeliski, R. (2011). Computer vision: Applications. Springer Science & Business Media

.

Howse, J. (2020). Learning OpenCV 4 computer vision with Python. Packt Publishing Ltd.