Extensive CNN Experiments for Action Recognition on ResActionsImages Variants and Transfer Learning

Introduction

In this assignment, we reconstruct and build upon the method presented in the conference paper *"An Image Representation of Skeletal Data for Action Recognition Using Convolutional Neural Networks."* [1]. Our objective is to take the CNN model from the paper and conduct experiments to improve it, exploring ways to enhance its performance in action recognition on a specific dataset.

For this purpose, we will use the **PKU-MMD dataset**, a large-scale dataset specifically designed for human action recognition. It consists of extensive training video and depth sequences, offering rich motion data for deep learning-based approaches. We are given a dataset of pre-generated pseudo-colored images derived from 3D skeleton trajectories, which encode skeletal motions into visual representations.

Methodology

The **PKU-MMD dataset** contains instances recorded from three camera angles: M (Middle), R (Right), and L (Left). Since our goal is to evaluate the robustness and generalization of the proposed approach, we conduct experiments across multiple training and testing configurations:

- **M/M, R/R, L/L** – Training and testing on the same camera view.
- **MR/L, ML/R, RL/M** – Training on two camera views and testing on the third, to examine generalization across viewpoints.

The large-scale datasets consist of 11 classes, which are: *eat meal snack, falling, handshaking, hugging another person, make a phone call, answer phone, playing with phone/tablet, reading, sitting down, standing up, typing on a keyboard, and wear jacket.* Additionally, we experiment with a **smaller dataset**, which contains 15 classes: *wave right, swipe down right, swipe left, swipe up right, wrist rotation, swipe up left, clapping, zoom in, swipe right, hands squeeze, hands circle, swipe diagonal, swipe down left, zoom out, hand raise.*

The large-scale datasets contain 1,500 images each, while the small dataset consists of 300 images. All the datasets allow us to analyze model performance across different action classes, viewpoint changes, and dataset sizes, providing a comprehensive evaluation of the method.

In our experiments, we systematically explore various strategies to enhance the performance of the CNN model. We implement **L2 regularization** and introduce **dropout layers** to prevent overfitting, and experiment with different **hidden layer sizes, optimizers, learning rates, and activation functions**. Additionally, we apply **cross-validation, dataset balancing, and data augmentation** to improve generalization. At each stage, we keep the best configuration and use it as the starting point for the next experiments. Once the optimal model is determined, we evaluate it on the test set for each of the seven datasets. Furthermore, we apply **transfer learning** to the **small dataset**, experimenting with two approaches: fine-tuning only the classifier and fine-tuning the entire model to evaluate the effectiveness of pre-trained representations in action recognition. In the provided notebooks, you'll find a comprehensive explanation of our decisions and implementations with detailed comments and references.

Results

The results of our approach are summarized in below **table**.

Experiment	Train	Test	proposed
Single-view	M	M	86.39
	R	R	84.27
	L	L	82.64
Cross-view	MR	L	87.56
	ML	R	91.36
	RL	M	93.48
Single-view	Small	Small	80.00
Single-view (TL)	Small	Small	83.33

Conclusions

Our experiments show that the applied enhancements to the proposed method effectively recognize human actions from 3D skeletal trajectories transformed into pseudo-colored images. Throughout our experiments, we tested different models, consistently selecting the best-performing configuration for final evaluation. The classification results indicate that our models achieved high accuracy, even in cross-view scenarios, where training and testing data come from different camera angles. We observed the effectiveness of our experimental workflow, such as L2 regularization, dropout, cross-validation and data augmentation, as well as the transformative impact of data, evidenced by higher results from combined datasets, and the considerable benefits of transfer learning in boosting overall performance. The promising results validate the effectiveness of our approach, highlighting a reliable methodology for performing such tasks efficiency.

Potential improvements include exploring different architectures, applying one of these models to real-world scenarios, and evaluating these applications using **GANs** to evaluate how effectively CNNs can be fooled by adversarial examples.

References

[1] Vernikos I., Mathe E., Papadakis A., Spyrou E., Mylonas P. (2019). *Image Representation of Skeletal Data for Action Recognition Using Convolutional Neural Networks.*