CSC 8850 - Advanced Machine Learning
Project Proposal

Name	Email
Donggeun Yoo	dyoo12@student.gsu.edu
Duc Quan Do	ddo17@student.gsu.edu

Human Action Classification

1. Introduction

Human Action Recognition (HAR) is a fundamental challenge in computer vision with diverse applications, including automated surveillance, human-computer interaction, and healthcare monitoring. The primary goal is to accurately classify human movements from video sequences, which involves interpreting both spatial appearance and temporal dynamics.

Despite recent advancements in deep learning, classical machine learning approaches remain highly relevant for understanding the core principles of spatiotemporal feature representation and model efficiency. This project aims to implement a robust classification pipeline using the benchmark KTH Action Recognition dataset (Schuldt et al., 2004). By comparing a baseline model with more complex classifiers like SVM or Bayesian classifiers, we evaluate how different statistical learning methods handle the complexities of motion and intra-class variations across different environmental scenarios.

2. Dataset Overview

The KTH Action Recognition dataset, which is available at https://www.kaggle.com/datasets/vafaeii/kth-action-recognition-dataset/data, is one of the most widely used benchmarks for human activity analysis. It provides a structured environment to test the robustness of classifiers against variations in appearance and surroundings.

- Action Categories: The dataset contains six types of human actions: walking, jogging, running, boxing, hand waving, and hand clapping.
- Scenarios: To simulate real-world conditions, these actions are performed by 25 different subjects across four distinct scenarios:
    - s1: Outdoors
    - s2: Outdoors with scale variations
    - s3: Outdoors with different clothes
    - s4: Indoors
- Data Scale: The dataset consists of 2,391 video sequences. Each video is captured at a frame rate of 25 fps with a resolution of 160x120 pixels.
- Characteristics: All videos are recorded against relatively static backgrounds, allowing the models to focus on the movement of the subjects. However, the variations in lighting (indoor vs. outdoor) and clothing provide a significant challenge for achieving generalization.

3. Methodology

Our process follows a standard classification pipeline consisting of video preprocessing, feature extraction, model training, and evaluation.

First, we will decode each video from the original AVI format and convert it into a sequence of greyscale frames to reduce the dimensionality and computational cost. Then, for each sequence, we will perform uniform sampling to a fixed number of frames to make sure the length is the same for all sequences. From the processed frames, we will extract spatiotemporal features to represent both appearance and motion information, which are aggregated across time to form feature vectors for the classifiers.

To ensure subject-independent evaluation and prevent identity leakage, we adopt the subject-based date split proposed in the KTH Action Recognition study (Schuldt et al., 2004), with separated subjects assigned to the training (70%), validation (15%), and testing (15%) sets. We will use the validation set to select the hyperparameters, and use the test set to report the final results.

We will implement three classifiers to study the trade-off between model assumptions and classification performance. We choose Support Vector Machine to be the baseline model for this study as it provides a direct point of comparison with prior work conducted by Schuldt et al. (2004). For an alternative probabilistic approach, we will implement a Bayesian classifier, which is theoretically optimal in minimizing classification error under correct model assumptions. In addition, we will include a distance-based method that does not estimate probability density functions, such as k-Nearest Neighbors or Minimum Distance Classifier, to examine how simpler decision rules perform on this task.

Given the multi-class nature of the problem, model performance is evaluated using balanced accuracy (BA), macro-averaged F1 score, and confusion matrices. Balanced accuracy accounts for prediction performance, the F1 score reflects robustness through precision and recall, and confusion matrices provide insight into common misclassification patterns among action classes.

4. Tentative Timeline

Date	Task Description
02/01	Finalize the model selection
02/08	Finish implementing the baseline model
02/28	Finish implementing other models
03/09	Analyze the results and brainstorm ideas on how to improve the performances
03/12	Finish improvement
03/18	Finalize project report and presentation slides

References

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: A local SVM approach. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* https://doi.org/10.1109/icpr.2004.1334462