Optical Flow Calculation, Embedding, and Classification for Action Recognition

Mikel Ballay
mikel.ballay@ufl.edu

Jaime Lobato
jaime.lobato@ufl.edu

December 3, 2025

Abstract

Action recognition plays a crucial role in applications like video surveillance, human-computer interaction, and sports analytics. This paper investigates the use of optical flow for action recognition, encompassing its calculation, embedding, and classification. Leveraging a pre-trained BN-Inception model with temporal segment networks (TSN) for optical flow embeddings and a BERT-based architecture for classification, the approach achieves notable results on the ECOLE dataset. Despite challenges, the integration of motion dynamics proves promising. Future work aims to combine optical flow with RGB features and test alternative datasets.
Keywords: *action recognition ; optical flow ; feature embedding ; classification .*

1 Introduction

Action recognition is integral to various domains, including video surveillance and sports analytics. A motion-centric approach like optical flow is a compelling choice, focusing on motion dynamics while ignoring extraneous features like color. This study explores the full pipeline: from optical flow extraction to feature embedding and action classification.

The primary objectives are:

1. **Calculation**: Extract optical flow patterns representing motion.
2. **Embedding**: Compactly represent these patterns for efficient processing.
3. **Classification**: Use a sequence model to categorize actions based on motion patterns.

This project employs a pre-trained BN-Inception model for feature extraction and a custom BERT-based transformer for classification. The ECOLE dataset serves as the evaluation benchmark.

?

2 Methodology

2.1 Optical Flow Calculation

Optical flow captures motion between consecutive video frames. The process:

- **Grayscale Conversion**: Convert frames to grayscale to isolate motion.
- **Pixel Displacement**: Compare pixels to calculate motion vectors.

- **Visualization**: Represent motion using directional arrows or colors for intensity.

2.2 Feature Embedding

We adapt BN-Inception, pre-trained on TSN, to extract 10-channel optical flow embeddings. Modifications include:

- **Input Layer**: Configured for stacked optical flow frames.
- **Output Layer**: Modified to output embeddings instead of classification scores.

2.3 Classification

A transformer-based BERT model processes the embeddings. Adaptations include:

- **Frame Tokenization**: Frames are treated as tokens, with position embeddings for sequence order.
- **Self-Attention**: Captures temporal relationships in motion patterns.
- **Classification Token**: Aggregates sequence information for final decision-making.

Training optimizations, including padding and normalization, enhance efficiency.

(**?**)

3 Results

Final Results on ECOLE:

- Accuracy: 15.72%

These results reflect the potential and challenges of using optical flow alone for action recognition.

Per-Class accuracy:

- 'swing': 9.99%
- 'unplug': 5.00%
- 'wave': 33.33%
- 'dance': 33.33%
- 'ignite': 5.00%
- 'plug': 15.00%
- 'squat': 57.14%
- 'pick_up': 45.00%

- 'detach': 13.64%
- 'roll': 12.24%
- 'scrap': 35.00%
- 'bend': 20.00%
- 'put_down': 55.00%

Classes with 0 percent accuracy: 'kick', 'flip', 'zibacoon', 'filter', 'slide', 'deflate', 'hue', 'skewer', 'catch', 'pour', 'stir', 'attach', 'hammer', 'push', 'run', 'hide', 'nom', 'fold', 'jump', 'cut', 'crawl', 'walk', 'climb', 'twist', 'dip', 'talk', 'throw', 'carve', 'chop', 'pull', 'duck', 'inflate', 'convex'.

Conclusions from per-class accuracy:

The per-class accuracy results highlight significant variability in the performance of optical flow-based classification, with some classes achieving relatively higher accuracy (e.g., 'squat': 57.14%, 'pick_up': 45.00%, 'put_down': 55.00%) while others consistently score 0%. This disparity can be attributed to the nature of the motion patterns associated with each class. Optical flow captures changes in motion over time, making it more effective for actions characterized by distinctive, large-scale, and consistent movements, such as 'squat' or 'wave.' These actions generate clear and discernible motion patterns, which are easier for the model to identify and classify.

In contrast, classes with subtle, ambiguous, or highly variable movements (e.g., 'hide,' 'kick,' or 'talk') struggle to achieve meaningful accuracy. Optical flow alone cannot adequately capture the spatial or contextual information required to distinguish such actions, as it primarily focuses on motion rather than appearance or object-related cues. Furthermore, the lack of temporal consistency in some actions may further degrade performance. This underscores the limitation of using optical flow as the sole feature for action recognition and suggests that integrating additional modalities, such as RGB appearance data or depth information, would likely enhance classification accuracy across all classes.

4 Discussion and Conclusions

Challenges faced included the computational burden of processing the ECOLE dataset, difficulties in replicating benchmarks, and the inherent limitations of optical flow. Our findings suggest that optical flow alone is insufficient for robust action recognition.

Future Directions:

1. Combine optical flow with RGB and depth features.
2. Test the model on HMDB and UCF101 datasets.
3. Optimize hyperparameters to improve performance.

The integration of diverse modalities and datasets promises to significantly advance action recognition models.