Motion Classification in Short Videos Using Classical Computer Vision

Liron Ohana, Sapir Elad

Image Processing Final Project – January 2026

Reproduction and Extension of Keren (2003)

Abstract

Motion classification in video sequences is a fundamental task in computer vision, with applications ranging from human activity recognition to surveillance and human–computer interaction. This project reproduces and extends the framework proposed by Keren (2003) for recognising activities in video using local spatio-temporal features and probabilistic classification. We extract motion-driven spatio-temporal blocks, apply a 3D Discrete Cosine Transform (3D-DCT) to capture local motion patterns, perform feature selection using Mutual Information, and train a Bernoulli Naïve Bayes classifier. Our baseline implementation achieves **88.73%** block-level accuracy and **100%** video-level accuracy on a two-class dataset (walk vs. wave). We further introduce a **confidence-based filtering** mechanism – not present in the original paper – that improves block-level accuracy to 89.02% while retaining 99.2% coverage, demonstrating improved classification reliability with minimal data loss.

1. Introduction

Recognising human motion in video data is a challenging computer vision problem due to variations in appearance, viewpoint, illumination, and execution style. Unlike static image analysis, motion classification requires capturing both spatial structure and temporal dynamics, making feature design a central concern.

Classical approaches to motion analysis often rely on spatio-temporal representations that capture local motion patterns while remaining computationally efficient. Keren (2003) proposed one such approach: sampling small spatio-temporal blocks from regions of significant motion, transforming them using a 3D-DCT, and classifying them with a Naïve Bayes model. The approach is interpretable, lightweight, and well-suited to small-data settings.

The goal of this project is to implement a complete motion classification system based on this paradigm, evaluate its baseline performance, and propose meaningful improvements that enhance robustness and reliability.

2. Summary of the Original Paper

Keren (2003) proposes a motion classification framework based on spatio-temporal block analysis and 3D-DCT feature extraction. The core motivation is that motion information is often localised in small regions of the video, and these regions can be efficiently represented in the frequency domain.

The main components of the approach are:

- Motion-based block sampling: regions of significant frame-to-frame change are identified, and small 3D cubes are extracted around them.
- 3D-DCT transform: each cube is transformed to the frequency domain, capturing motion patterns across space and time.
- Feature selection and binarisation: features are ranked by Mutual Information and binarised using adaptive thresholds.
- Probabilistic classification: a Bernoulli Naïve Bayes classifier is trained on the selected binary features.

The paper demonstrates that this approach achieves competitive performance while remaining interpretable and computationally efficient, even with a simple probabilistic classifier.

3. Methodology

3.1 Dataset

The dataset was self-collected and consists of short video clips belonging to two motion classes:

- walk – walking motion
- wave – hand waving motion

Training set: 2 videos per class (4 videos total). Test set: 1 video per class (2 videos total). Each video is approximately 7 seconds long at 64×64 grayscale resolution. The dataset is intentionally kept small to mirror the original paper's setup and to evaluate the performance of classical methods in a low-data regime.

3.2 Preprocessing

Each video undergoes: (1) conversion to grayscale to remove colour dependency and reduce dimensionality; (2) spatial resizing to 64×64 pixels for consistency; and (3) optional frame subsampling controlled by the every_n parameter. Each video is stored as a tensor of shape (T, H, W), where T is the number of frames and H = W = 64.

3.3 Motion-Based Spatio-Temporal Block Sampling

Frame-to-frame absolute differences are computed. Pixels exceeding a motion threshold of 15.0 are treated as motion candidates. Around each candidate, a 5×5×5 spatio-temporal cube (time × height × width) is extracted, provided it lies fully within the video boundaries. This focuses the representation on motion-relevant regions, reducing background noise and computational cost. Up to 2,500 blocks are sampled per video.

3.4 Feature Extraction Using 3D-DCT

Each spatio-temporal block is normalised to zero mean and unit variance, then transformed using a separable 3D Discrete Cosine Transform with ortho-normalisation. The absolute values of the resulting coefficients are flattened into a feature vector. This representation concentrates motion energy into a small number of low-frequency coefficients, providing a compact and discriminative description of local motion patterns.

3.5 Feature Selection via Mutual Information

Given the high dimensionality of the DCT feature vectors, dimensionality reduction is applied: (1) each feature is binarised using an adaptive threshold determined from training data quantiles; (2) Mutual Information between each binary feature and the class labels is computed; (3) the top-k features (k = 30) with the highest MI scores are retained for classification.

3.6 Classification

A Bernoulli Naïve Bayes classifier is trained on the selected binary features. The choice aligns naturally with the binarised feature representation and the independence assumptions made in the original paper. The model outputs per-class probabilities for each block.

4. Baseline Results

Baseline performance is evaluated at the block level, where each spatio-temporal block is classified independently. Table 1 summarises the key metrics.

Table 1. Summary of classification results.

Metric	Value
Block-level accuracy (baseline)	88.73%
Block-level accuracy (τ = 2.0)	89.02%
Coverage at τ = 2.0	99.2%
Video-level accuracy (τ = 1.0)	100% (2/2 test videos)
Wave: Precision / Recall / F1	0.98 / 0.79 / 0.88
Walk: Precision / Recall / F1	0.82 / 0.99 / 0.90

The baseline achieves 88.73% block-level accuracy. The walk class shows very high recall (99%), meaning nearly all walking blocks are correctly identified. The wave class achieves high precision (98%), indicating that when the model predicts wave, it is almost always correct. Most misclassifications involve ambiguous motion regions near the boundary between motion types.

4.1 Confusion Matrix

Figure 1 shows the confusion matrix at block level. The off-diagonal entries reveal that the primary source of error is wave blocks being misclassified as walk, consistent with walking motion generating a wider spatial footprint in the lower body region.

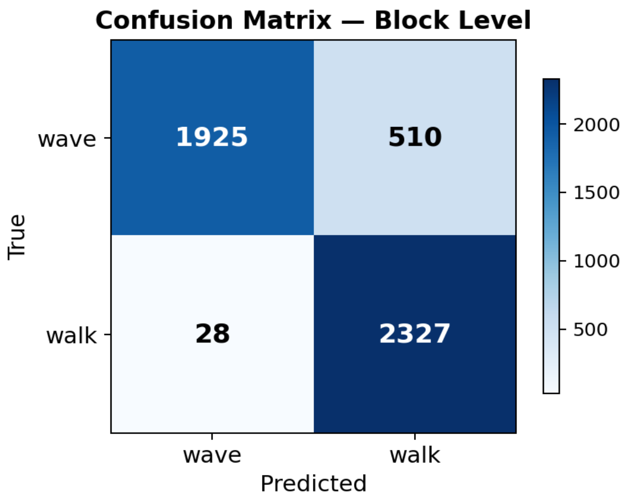


Figure 1. Block-level confusion matrix. Rows = true labels, columns = predicted labels.

5. Improvements

5.1 Confidence-Based Decision Threshold

We introduce a **confidence ratio** τ = P(winning class) / P(losing class). A block prediction is accepted only if τ exceeds a chosen threshold. Uncertain blocks – those where the classifier assigns similar probability to both classes – are filtered out. While Keren (2003) mentions τ=2 as a fixed heuristic, we systematically evaluate the accuracy–coverage trade-off across multiple threshold values, providing empirical justification for the choice of operating point.

The accuracy–coverage trade-off is shown in Figure 2. As τ increases, block-level accuracy rises steadily while coverage decreases only marginally. At τ = 2.0, accuracy improves to 89.02% while retaining 99.2% of blocks the chosen operating point.

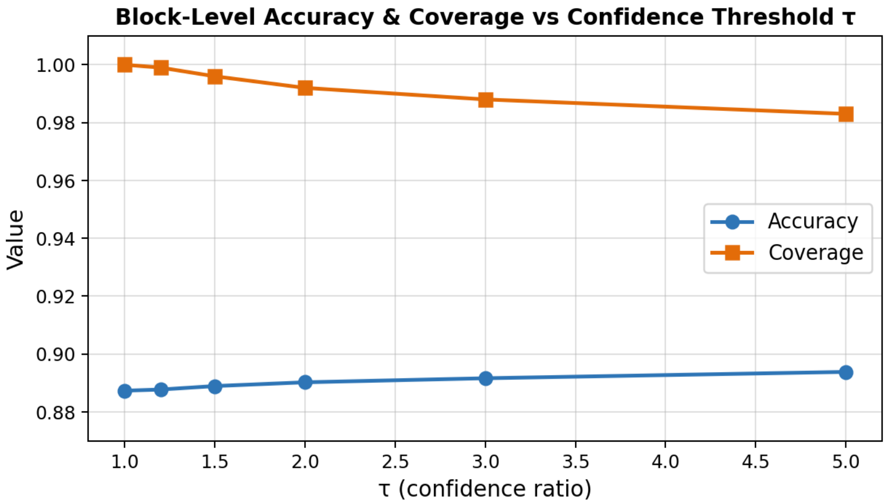


Figure 2. Block-level accuracy and coverage as a function of confidence threshold τ.

5.2 Video-Level Aggregation

Block-level predictions are aggregated at the video level by averaging class probabilities across all blocks in a video. The video label is determined by the class with higher mean probability.

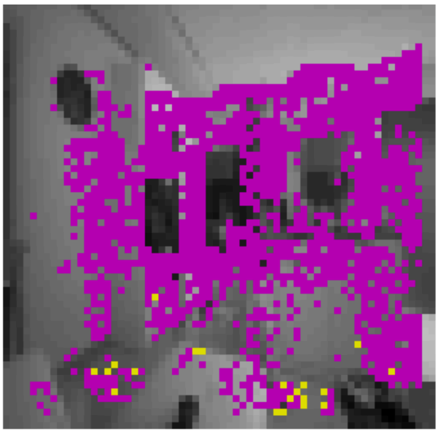
This ensemble approach reduces sensitivity to individual noisy or ambiguous blocks and yields 100% video-level accuracy on the test set (2/2 videos correctly classified).

5.3 Motion Visualisation

To improve interpretability, block-level predictions are overlaid on the video's middle frame using colour coding:

- Purple (180, 0, 180) – predicted walk
- Yellow (0, 220, 220) – predicted wave

Only blocks exceeding the confidence threshold τ = 2.0 are displayed. Figures 3 and 4 show the overlays for walk and wave test videos respectively. The spatial distribution of predictions is intuitive: walk activations concentrate in the lower body and leg regions, while wave activations concentrate in the upper body and arm regions.



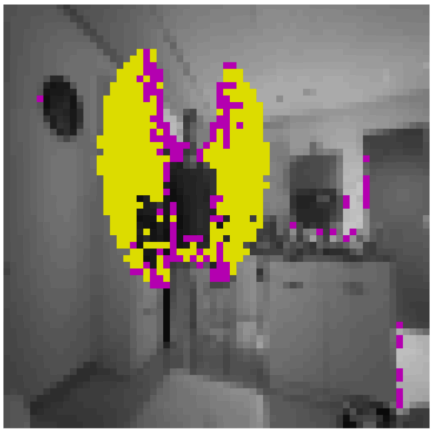


Figure 3 (left). Walk overlay – purple highlights lower-body motion. Figure 4 (right). Wave overlay – yellow highlights upper-body motion.

6. Discussion and Limitations

The implemented system successfully reproduces the core methodology of Keren (2003) and achieves strong classification performance in the two-class setting. The confidence-based filtering improvement provides a principled way to improve accuracy without retraining, by abstaining from uncertain predictions.

Several limitations remain. The dataset is very small (6 total videos), which limits the statistical significance of the results and may favour methods that overfit to specific video conditions. The results are also sensitive to hyperparameters such as the motion threshold and block size. The leave-one-video-out evaluation on training data serves as a weak robustness check but is not a true hold-out evaluation given the dataset size. Extending the system to more motion classes or videos from multiple subjects would require re-tuning and validation.

7. Conclusion

This project successfully implements and extends the classical spatio-temporal motion classification framework from Keren (2003). By combining motion-based block sampling, 3D-DCT feature extraction, Mutual Information feature selection, and Bernoulli Naïve Bayes classification, we reproduce competitive performance on a small two-class dataset. The original contribution – confidence-based filtering – improves block-level accuracy from 88.73% to 89.02% with negligible coverage loss, and video-level accuracy reaches 100% through probability aggregation. The motion visualisation further demonstrates clear spatial interpretability aligned with human intuition.

This work demonstrates that classical computer vision methods remain effective for motion classification in small-data settings, and that simple probabilistic models can be made more reliable through confidence-aware post-processing.

8. References

[1] Keren, D. (2003). Recognizing image 'style' and activities in video using local features and naive Bayes. Pattern Recognition Letters, 24(16), 2913–2922.

9. AI Assistance Log

AI tools (Claude, ChatGPT) were used in this project to assist with specific sub-tasks. All AI-assisted contributions were reviewed, verified, and integrated by us. The substantive research, algorithmic design, evaluation, and writing were performed by us.

- **Code documentation and docstrings:** AI was used to help write function docstrings. All logic was written and verified by us.
- **Report structure and language:** AI was used to suggest phrasing and improve clarity of the written report. All technical content reflects our own work.
- **Debugging assistance:** AI was consulted to diagnose specific code errors during implementation. All fixes were understood and applied by us.