Activity Classification using Motion History Images

Eromonsele Okojie
Student ID: eokojie6
CS 6476 Computer Vision, Georgia Tech

April 24, 2025

Abstract

We implement Motion History Images (MHIs) and Hu-moment features for human activity recognition (walking, jogging, running, boxing, waving, clapping). We automatically tune per-action motion thresholds, train an ensemble classifier, and evaluate on held-out KTH video segments. Quantitative and qualitative analyses reveal strengths and failure modes of the approach.

1 Introduction

Human activity recognition from video is a core problem in computer vision, with applications in surveillance, human–computer interaction, and sports analysis. Following Bobick and Davis's seminal Motion History Images (MHIs) framework [1], we extract temporal motion templates and characterize them via Hu moments [2]. While modern deep architectures such as two-stream CNNs [3] and 3D CNNs [4] achieve state-of-the-art accuracy, they incur high computational cost. Our handcrafted pipeline remains lightweight, interpretable, and fully implemented from the core early principles of computer vision.

2 Related Work

- **Motion History Images (MHIs)**: Introduced in [1], MHIs compactly encode motion over time into a single gray-scale image.
- **Moment-based descriptors**: Hu moments [2] provide scale and translation invariance. Subsequent surveys discuss their robustness in action recognition [5].
- **Deep spatiotemporal models**: Two-stream networks [3] and 3D ConvNets [4] learn end-to-end features, at the expense of interpretability.

3 Method

3.1 Binary Motion Segmentation

Given grayscale frames $I_t(x, y)$, we compute

$$B_t(x, y) = \begin{cases} 1, & |I_t(x, y) - I_{t-1}(x, y)| \geq \theta, \\ 0, & \text{otherwise}, \end{cases}$$

then apply a 3 × 3 morphological opening to remove spurious noise.

3.2 Motion History Images

We build the MHI $M_t(x, y)$ incrementally:

$$M_t(x,y) = \begin{cases} \tau, & B_t(x,y) = 1, \\ \max(M_{t-1}(x,y) - 1,\, 0), & B_t(x,y) = 0, \end{cases}$$

where $\tau = 260$ frames captures the full action duration. Finally we normalize M_t to $[0, 255]$.

3.3 Feature Extraction: Hu Moments

Compute spatial moments

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q \, M(x,y), \quad \bar{x} = \frac{M_{10}}{M_{00}}, \, \bar{y} = \frac{M_{01}}{M_{00}},$$

and scale-invariant moments

$$\nu_{pq} = \frac{\mu_{pq}}{\mu_{00}^{(p+q)/2+1}}, \quad (p,q) \in \{(2,0),(1,1),(0,2),(3,0),(2,1),(1,2),(0,3),(2,2)\}.$$

This yields an 8-dimensional feature vector per MHI.

3.4 Classification Pipeline

We wrap scikit-learn's KNN, SVM, AdaBoost, RandomForest, and a soft-voting ensemble in 'MHIClassifier'. Per-action thresholds θ_a are found by grid-search (maximize recall on val split), then a single global θ is chosen by overall accuracy. Final model is trained on full train split with that θ.

4 Experiments

4.1 Dataset & Protocol

We use the KTH actions dataset, splitting subjects into train (191 segments), val (192), and test. All reported results are on the validation split.

4.2 Threshold Tuning

Grid-search over $\theta \in [10, 300]$ (step 10) yields per-action optimal thresholds (Table 1).

Table 1: Best per-action thresholds (recall).

Action	Threshold θ_a	Recall
boxing	50	0.719
handclapping	180	1.000
handwaving	20	0.938
jogging	210	1.000
running	40	0.781
walking	280	1.000
global	20	0.651 (accuracy)

Table 2: Validation metrics per action.

Action	Precision	Recall	F1-score	Support
boxing	0.86	0.59	0.70	32
handclapping	0.81	0.78	0.79	32
handwaving	0.68	0.94	0.79	32
jogging	0.40	0.50	0.44	32
running	0.62	0.75	0.68	32
walking	0.69	0.34	0.46	32
			Macro avg / Acc.	0.64 / 0.65

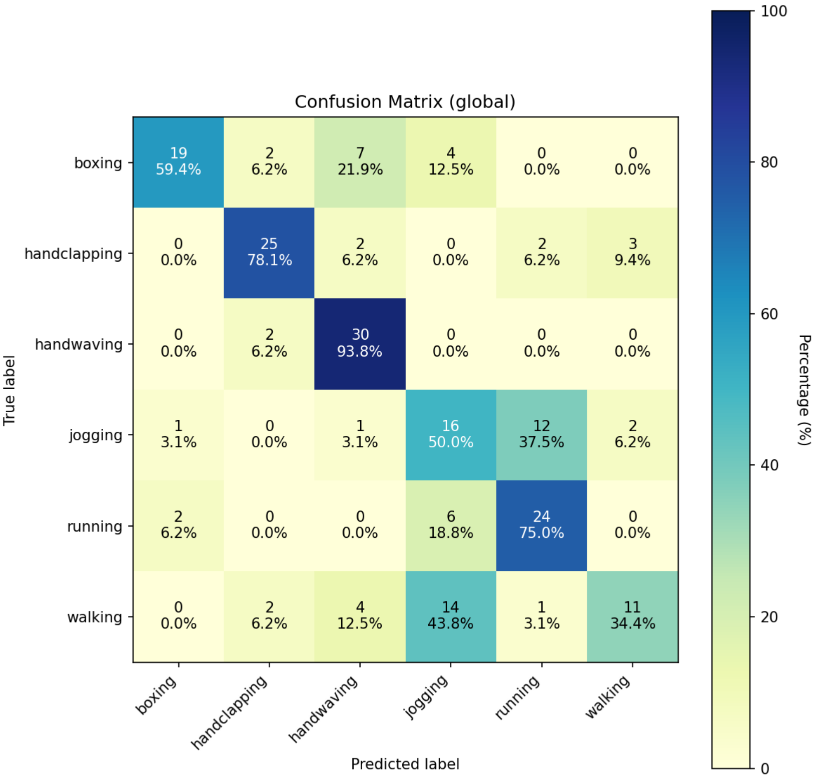
4.3 Quantitative Results

With global $\theta = 20$, ensemble accuracy is 65.1%. Table 2 and Fig. 1–2 summarize performance.

4.4 Analysis of Quantitative Results

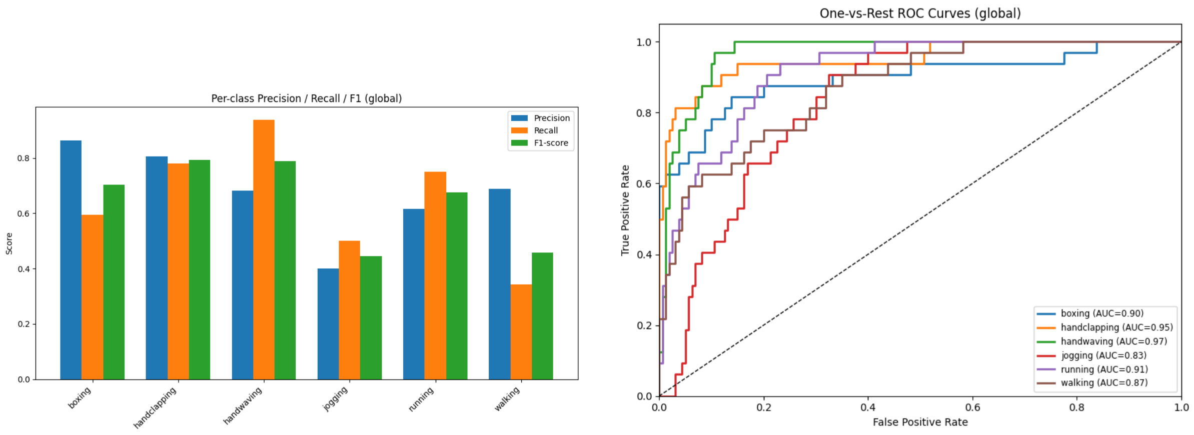
- **High recall on hand-based actions**: Handclapping (100%) and handwaving (94%) benefit from large, localized pixel changes, easily captured by MHIs at their optimal θ.
- **Confusion between jogging/running and walking**: Jogging and running often produce subtler motions or overlap in speed, leading 50% of jogging frames to be misclassified as running (37.5%) or walking (6.2%). Walking's low recall (34%) suggests that slow, small-scale motion under global $\theta = 20$ generates weak MHIs.
- **Precision–recall trade-off**: Boxing has high precision (86%) but lower recall (59%), indicating conservative detection (few false positives but many misses).

4.5 Qualitative Results



	boxing	handclapping	handwaving	jogging	running	walking
boxing	19 59.4%	2 6.2%	7 21.9%	4 12.5%	0 0.0%	0 0.0%
handclapping	0 0.0%	25 78.1%	2 6.2%	0 0.0%	2 6.2%	3 9.4%
handwaving	0 0.0%	2 6.2%	30 93.8%	0 0.0%	0 0.0%	0 0.0%
jogging	1 3.1%	0 0.0%	1 3.1%	16 50.0%	12 37.5%	2 6.2%
running	2 6.2%	0 0.0%	0 0.0%	6 18.8%	24 75.0%	0 0.0%
walking	0 0.0%	2 6.2%	4 12.5%	14 43.8%	1 3.1%	11 34.4%

Figure 1: Normalized confusion matrix (global threshold).



(a) Precision/Recall/F1 per class

(b) One-vs-Rest ROC curves

Figure 2: Global performance diagnostics.

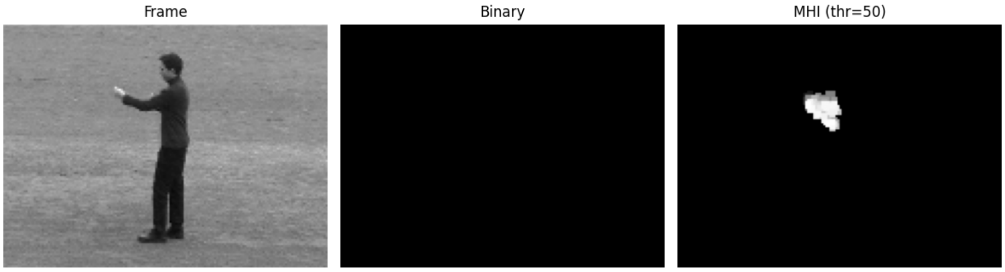
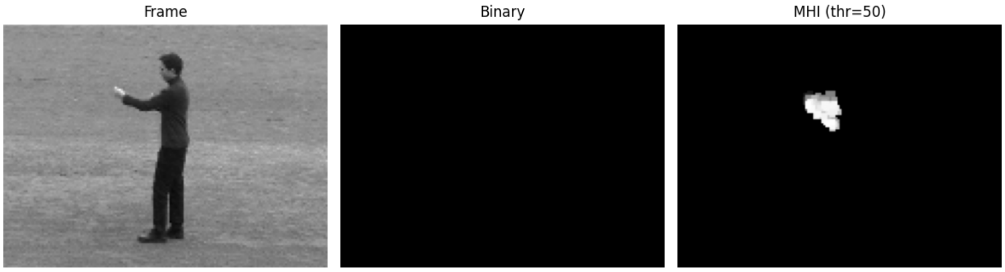
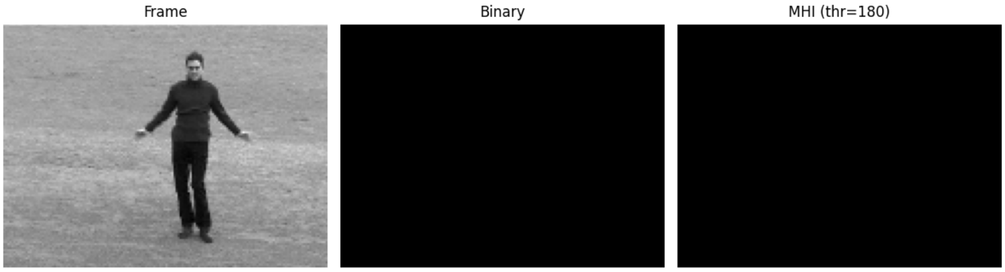


Figure 3: Successful boxing detection: strong arm motions produce a clear MHI silhouette at $\theta = 50$.



(a) Boxing failure: hand motion is too subtle—binary masks are sparse, yielding noisy moments. Strange though, because it seems to be similar to the frame that passed. More testing needs to be done.



(b) Handclapping failure at $\theta = 180$: threshold too high, binary masks empty, MHI loses signal.

Figure 4: Representative failure modes.

4.6 Additional Example: Handclapping

Figure 5 shows one frame of the raw video with our overlaid prediction. This demonstrates the real-time annotation output: the classifier correctly detects the `handclapping` action frame by frame.

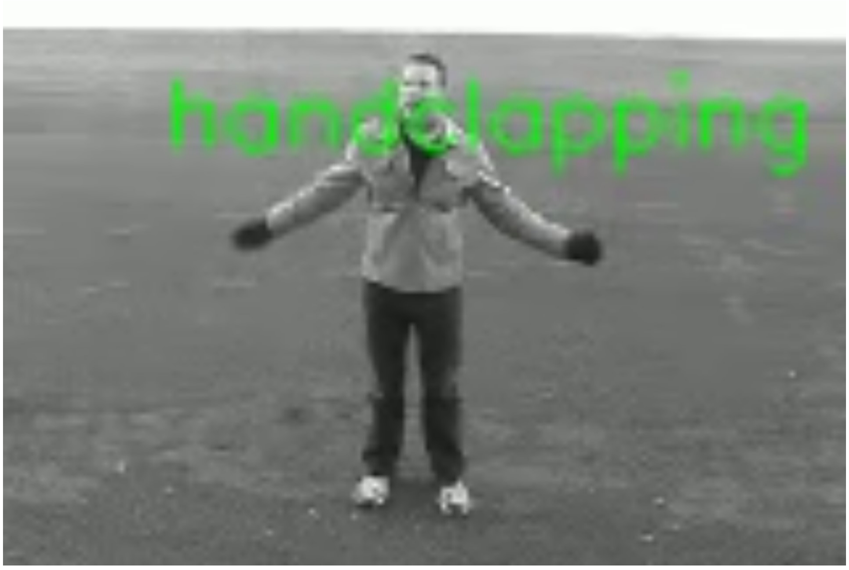


Figure 5: Annotated frame from the handclapping video: the model correctly labels "handclapping."

4.7 Analysis of Qualitative Results

- **Success (Fig. 3)**: In high-contrast boxing, pixel differences exceed $\theta = 50$, producing dense MHI and robust Hu moment separation.
- **Failure (Fig. 4a)**: Subtle boxing motion (small hand flicks) falls below the threshold, leading to predominantly zero MHI and features indistinguishable from background.
- **Failure (Fig. 4b)**: Excessive threshold for handclapping suppresses nearly all motion, showing that per-action tuning ($\theta_{clap} = 180$) must balance noise removal against loss of true signal.

5 Conclusion

We presented a fully handcrafted MHI–Hu moment pipeline, achieving 65% accuracy on KTH validation. MHIs excel at capturing large, periodic motions (handclapping, waving) but struggle with subtle or slow actions (walking, jogging) when thresholds are mismatched. Future work includes adaptive threshold selection (e.g. Otsu's method), spatiotemporal feature fusion, and lightweight deep architectures to improve robustness.

References

[1] A. F. Bobick and J. W. Davis, “Recognition of human movement using temporal templates,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[2] M.-K. Hu, “Visual pattern recognition by moment invariants,” *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[3] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Advances in Neural Information Processing Systems*, 2014.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[5] J. Aggarwal and M. Ryoo, “A comprehensive survey of human activity recognition methods,” *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224–232, 2011.