Action and Activity Recognition

Exploring Motion History Images for Human Action Classification

By: Khalid Mahmoud, Ethan Grosvenor, Seth Ojo, Devstutya Pandey, Mack Bourne

What is Action and Activity Recognition?

- The main idea of action recognition is to be able to identify human movements from video data.
- Has a wide range of real-world applications:

  - Sports analytics for tracking player performance or techniques used.
  - Surveillance and security to improve threat detection
  - Healthcare & Rehabilitation – Monitoring patients and assisting therapy.
  - Human-Computer Interaction – Enabling gesture control in gaming and VR.

Traditional CV vs. Deep Learning: The Smarter Choice?

Challenges with Deep Learning:

- Requires massive datasets and **long training times**.
- Demands **high computational power (GPUs, TPUs)**.
- Acts as a **black box**, making decisions hard to interpret.

Advantages of Traditional Computer Vision:

- Works with **smaller datasets**.
- Runs efficiently on **standard CPUs**.
- Provides **interpretable features**, making debugging easier.
- **Better for real-time applications** with lower computational costs.

Motion History Image (MHI): Capturing Motion Over Time

What is MHI?

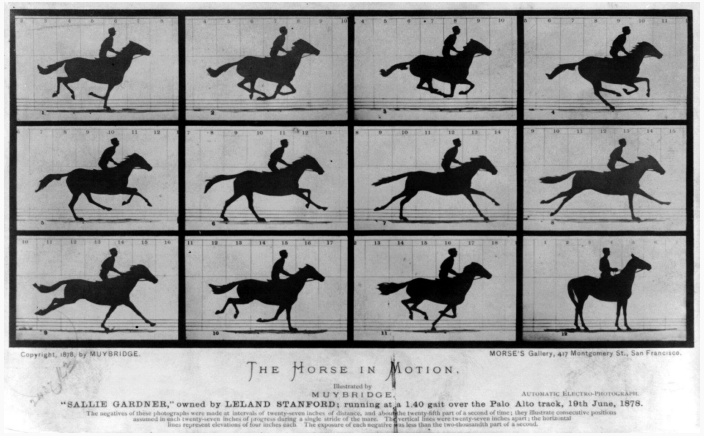
- A grayscale image that represents the recent history of motion in a scene.
- Brighter pixels indicate more recent movement, while darker pixels fade over time.

How It Works

- Constructed by updating pixel values based on motion detected in consecutive frames.
- Uses a temporal decay function to gradually fade older motion while emphasizing recent activity.

Why Use MHI?

Encodes motion timing: Differentiates between actions like walking vs. running.
Captures motion direction: Tracks movement flow by showing motion buildup.
Compact representation: Summarizes motion history in a single image.



Key Characteristics

- Works with **binary silhouette masks** to isolate motion.
- Suitable for **gesture recognition, action classification, and motion tracking**.
- Often paired with feature extraction methods like **Hu Moments or Projection Profiles**.

Challenges in Traditional MHI-based Action Recognition

Why Classic MHI Falls Short

No Motion Intensity Awareness – Treats all movements the same, even if speeds differ (**optical flow weighting fixes this**).
Loses Temporal Order – Captures only the final movement shape, not the full sequence.
No Directional Awareness – Cannot distinguish between left/right or up/down movements (**partially improved with optical flow**).
Limited to Static Backgrounds – Background movement can introduce noise, requiring preprocessed silhouette masks.

Why We Chose MHI for Our Project

Lightweight & Explainable – A simple, interpretable method without deep learning complexity.
Structured Processing Pipeline – Motion extraction → Temporal template (MHI) → Feature extraction → Classification.
Proven but Under-Optimized – Used in well-cited research but hasn't seen much recent innovation, making it a strong candidate for enhancements like optical flow integration.

Basis for Our Project

- The paper by Elit Cenk Alp and Hacer Yalim Keles (IEEE) presents a method for action recognition using Motion History Images (MHI) and Hu moments, combined with Hidden Markov Models (HMMs).
- This approach achieves 96% classification accuracy on the Weizmann dataset and serves as a strong foundation for action recognition systems.

How We Build On It

- Improved Feature Extraction: We extend their approach by adding advanced motion features like optical flow and gradient-based techniques for more precise motion representation.
- **Dynamic Action Representation**: Our use of Flow-weighted MHI and silhouette masks provides more detailed and dynamic motion data compared to the original MHI-based method.
- **Enhanced Temporal Understanding**: We improve temporal modeling through methods like temporal pyramids, which help track actions over time and improve overall accuracy.

Areas for Improvement

- Complex Backgrounds: While the paper shows robust results, real-world environments often contain more complex backgrounds and camera motion. We aim to enhance robustness in such scenarios.
- **Performance in Outdoor Settings**: The method performs well on controlled datasets but may struggle in more diverse or less structured outdoor settings, a challenge we address with better feature extraction techniques.

Making MHI Better: Our Innovations

- Optical Flow-Weighted MHI

Uses Optical Flow (e.g., Farneback) – Computes pixel-wise motion velocity between frames.

Improves Motion Intensity Representation – Differentiates between fast/slow movements, enhancing motion tracking.
Enhances Temporal Dynamics – Provides a more nuanced representation of movement speed and strength over time.
- **Temporal Pyramid MHI Representation**

Divides Video into Time Segments – Breaks down video into smaller time windows to analyze finer motion details.
Computes MHI/MEI for Each Segment – Treats each time segment independently for better motion understanding.

Captures Motion Evolution Over Time – Allows for the tracking of motion development and changes at different time scales.

- Directional Motion Energy Images (DMEI)

Separate Motion Masks for Each Direction – Identifies and isolates motion in the up/down and left/right directions.

Extracts HOG & LBP Features – Uses Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP) to capture motion features from each direction.
Improves Classification Accuracy – Better feature extraction allows for more accurate action recognition and classification.

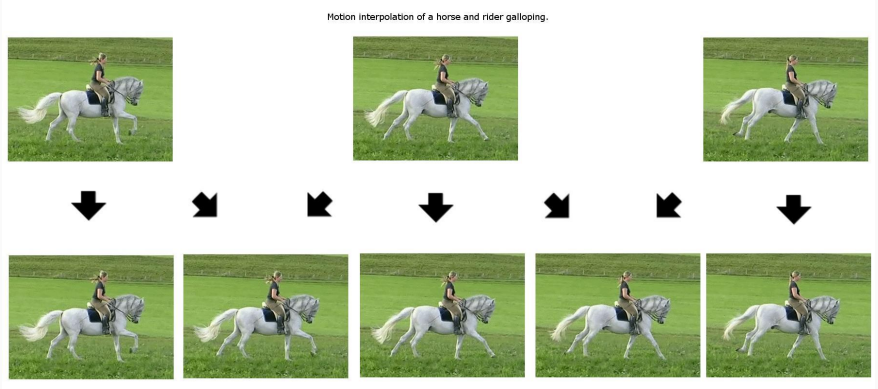
How Our Action Recognition System Works

1. Input Data

- **Reads Silhouette Masks**: We begin by processing silhouette video data. Each action (e.g., walking, running) is performed by different people, and the silhouette images are clean and pre-processed.
- **Multiple People, Same Actions**: Each action is performed by multiple individuals, ensuring the model can generalize across different people performing the same action.

2. Temporal Motion Encoding with MHI

- **Creating MHI**: We create Motion History Images (MHI) by comparing each frame to the previous one.
- **Brighter Pixels = Recent Motion**: MHI uses pixel brightness to represent the recency of motion. The more recent the motion, the brighter the pixel.
- **Tracking Movement Over Time**: This temporal decay process allows us to capture how motion changes over time, forming a visual representation of the action.



How our Action Recognition System Works (Continued)

3. Feature Extraction

- **Hu Moments**: We use Hu Moments, which are mathematical descriptors that capture the overall shape of the motion. These are good at summarizing the structure of human movement, like the general form of a person walking or running.
- **Projection Profiles (Optional)**: Alternatively, we can use projection profiles, which are computed by summing the pixel values in rows and columns of the MHI. This helps capture the motion distribution across the image.

4. Temporal Modeling with HMM

- **Hidden Markov Models (HMMs)**: We use HMMs to model how actions unfold over time. HMMs are great for understanding sequential patterns and can be trained on multiple examples of the same action.
- **Leave-One-Out Cross-Validation**: To ensure our system works well across different people, we train and test it by holding out one sequence at a time for each action, testing how well the system generalizes.
- **Improved Temporal Transitions**: We optionally use a structured approach to HMMs that ensures the model correctly tracks the order and flow of motion (left-to-right progression).

Comparison of New Method to Baseline

- **Baseline (Original MHI):**
  Accuracy: 97%
  Macro F1-score: 0.97
  Strengths: High performance across most action classes.
  Area for Improvement: Slight misclassification in waves (31 instances mistaken as side). Solid results but room for improvement in dynamic actions.

- **Our New Method (Optical Flow-Weighted MHI + Temporal Pyramid):**
  Accuracy: 98%
  Macro F1-score: 0.98
  Key Improvements:
    Stronger sensitivity to dynamic and subtle motions, especially in wave2 and side actions.
    F1-score for wave2 increased from 0.94 to 0.99, and for side from 0.90 to 0.97.
    Significant reduction in misclassifications for waves (from 36 to 6).

Performance Enhancements with Temporal Pyramid

Additional Improvements:

- **Jump Action:** F1-score improved from 0.95 to 1.00, showing better stability and precision.
- **Overall Stability:** Consistent performance across all action types, with higher recall for actions like jump (0.99+).
- **Robustness:** Improved overall prediction reliability, especially in handling actions with high variation.

Conclusion:

- Combining Optical Flow and Temporal Pyramid not only enhances accuracy but also improves stability and robustness in recognizing complex actions, making it a more balanced and reliable solution than the baseline method.

Baseline Data

Classification report:

	precision	recall	f1-score	support
bend	0.96	1.00	0.98	360
jack	1.00	1.00	1.00	450
jump	1.00	0.99	1.00	179
pjump	1.00	0.95	0.97	259
run	0.92	0.96	0.94	98
side	0.83	0.98	0.90	165
skip	0.97	0.98	0.97	176
walk	0.98	1.00	0.99	401
wave1	1.00	0.98	0.99	374
wave2	1.00	0.90	0.94	345
accuracy			0.97	2807
macro avg	0.97	0.97	0.97	2807
weighted avg	0.98	0.97	0.97	2807

First Improvement: Optical Flow-Weighted Motion History Image (MHI)

Classification report:

	precision	recall	f1-score	support
bend	0.99	0.99	0.99	360
jack	1.00	0.99	1.00	450
jump	0.92	0.98	0.95	179
pjump	1.00	0.96	0.98	259
run	0.99	0.97	0.98	98
side	0.96	0.98	0.97	165
skip	0.98	0.98	0.98	176
walk	0.96	1.00	0.98	401
wave1	1.00	0.98	0.99	374
wave2	1.00	0.98	0.99	345
accuracy			0.98	2807
macro avg	0.98	0.98	0.98	2807
weighted avg	0.98	0.98	0.98	2807

Second Improvement: Temporal Pyramid Motion Histogram Imaging

Classification report:

	precision	recall	f1-score	support
bend	0.66	0.80	0.72	369
jack	0.84	1.00	0.91	459
jump	0.95	0.74	0.83	188
pjump	0.85	0.81	0.83	268
run	1.00	0.96	0.98	107
side	1.00	0.87	0.93	174
skip	0.59	1.00	0.74	186
walk	0.99	1.00	1.00	411
wave1	0.74	0.53	0.62	383
wave2	0.73	0.51	0.60	354
accuracy			0.81	2899
macro avg	0.84	0.82	0.82	2899
weighted avg	0.82	0.81	0.80	2899

Combined optical flow weighted and temporal pyramid MHI

Classification report:

	precision	recall	f1-score	support
bend	0.96	1.00	0.98	360
jack	1.00	1.00	1.00	450
jump	1.00	0.99	1.00	179
pjump	1.00	0.95	0.97	259
run	0.92	0.96	0.94	98
side	0.83	0.98	0.90	165
skip	0.97	0.98	0.97	176
walk	0.98	1.00	0.99	401
wave1	1.00	0.98	0.99	374
wave2	1.00	0.90	0.94	345
accuracy			0.97	2807
macro avg	0.97	0.97	0.97	2807
weighted avg	0.98	0.97	0.97	2807

Improvement 3 that REDUCES accuracy: DHMI

Classification report:

	precision	recall	f1-score	support
bend	0.99	0.99	0.99	360
jack	1.00	0.99	1.00	450
jump	0.92	0.98	0.95	179
pjump	1.00	0.96	0.98	259
run	0.99	0.97	0.98	98
side	0.96	0.98	0.97	165
skip	0.98	0.98	0.98	176
walk	0.96	1.00	0.98	401
wave1	1.00	0.98	0.99	374
wave2	1.00	0.98	0.99	345
accuracy			0.98	2807
macro avg	0.98	0.98	0.98	2807
weighted avg	0.98	0.98	0.98	2807

Silhouette Masks

Silhouette masks are binary images that highlight the **outline of the subject** while ignoring the background and other irrelevant elements in a scene.

How They Help:
- In our project, **silhouette masks** allow us to focus on **human motion** without being distracted by the background or environmental details.
- By isolating the **subject's shape**, we can better analyze the changes in movement, enabling more accurate action recognition.

Advantages:
- Noise Reduction: Silhouettes reduce the noise from irrelevant background features, helping the model focus on the **core subject**.
- **Consistency:** Masks maintain **consistent representation** of the subject's shape, making it easier for the model to track motion over time.

Impact on Performance:
- The clean, pre-processed silhouette data ensures that the model can more accurately recognize complex actions and handle variations in subject appearance or background.

Temporal Motion Encoding - MHI

Creating MHI:

Motion History Images (MHI) are created by comparing **consecutive frames** in a video. This technique captures **temporal changes in motion**, helping us track how the subject's movement evolves over time.

Brighter Pixels = Recent Motion:

The pixels in an MHI image become **brighter** as the motion gets more recent. This makes it easier to visualize and track **recent actions** while differentiating between older and newer movements.

Tracking Movement:

The **temporal decay process** in MHI helps us track the progression of motion throughout the sequence. The MHI forms a **clear visual summary** of how the action unfolds, allowing the model to focus on both short-term and long-term motion patterns.

Feature Extraction - Hu Moments

Hu Moments:
Hu Moments are **mathematical descriptors** that capture the **overall shape and structure** of motion. They are used to summarize human movement patterns, such as walking or running, in a concise and effective manner.

Benefits:
Hu Moments are particularly **robust** in summarizing global motion features. They are **invariant** to changes in **scale, rotation**, and **translation**, making them ideal for recognizing and classifying motion regardless of variations in viewing angles or sizes.

Temporal Modeling - Hidden Markov Models (HMMs) & Validation

HMMs for Temporal Modeling:
Hidden Markov Models (HMMs) were used to model how actions unfold over time, capturing **sequential patterns** in motion data.
HMMs are ideal for recognizing **temporal patterns** and can be trained on **multiple sequences** of the same action to improve robustness.

Cross-Validation Approach:
We used **Leave-One-Out Cross-Validation** to train and test the model, holding out **one sequence** per action each time.

Generalization:
This method helped ensure that our system would **generalize well** to new individuals, improving its **robustness** and performance on **unseen data**.

Key Metrics

Accuracy:

- Accuracy measures how many of the predictions made by the system were correct. It's calculated by dividing the number of correct predictions by the total number of predictions.
- **How it contributes:** This is the most basic and widely used metric to see how well our system is performing overall. It helps us quickly understand how many of the actions are being correctly identified in the video data.

F1-score:

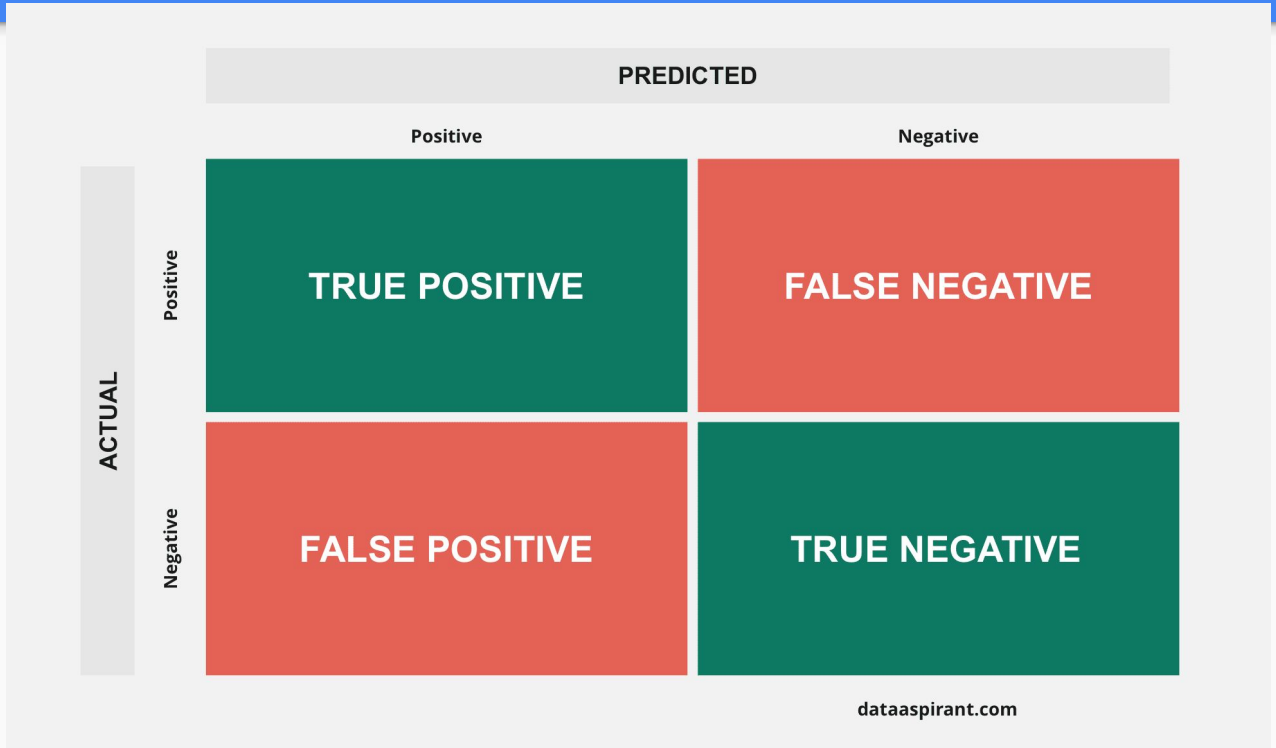
- The F1-score is the harmonic mean of precision and recall. Precision measures the accuracy of positive predictions, while recall measures how many actual positive cases were identified. The F1-score combines both into a single number, balancing precision and recall. It's especially useful when the dataset is imbalanced (e.g., some actions occur much more often than others).
- **How it contributes:** This metric is important when the action types are not evenly distributed. If one action is very common and another is rare, accuracy alone could be misleading. The F1-score gives a better sense of how well the model performs across all action types.

Confusion Matrix:

A confusion matrix is a table used to evaluate the performance of a classification algorithm. It breaks down the predictions into four categories:
- **True Positives (TP):** Correctly predicted actions.
- **False Positives (FP):** Incorrectly predicted actions.
- **True Negatives (TN):** Correctly identified non-actions.
- **False Negatives (FN):** Missed actions that should have been detected.

How it contributes: The confusion matrix helps visualize where the model is making errors and which action categories are more difficult to classify. It gives a deeper understanding of the system's performance beyond just accuracy.

Confusion Matrix



Baseline vs. Improved Methods Comparison

Baseline (Original MHI)
Accuracy: 97%
Key Observations: Strong performance across most action classes but some misclassification of similar actions (e.g., wave2 mistaken as side).
Limitations: Struggles with dynamic actions, particularly for subtle differences.

Optical Flow-Weighted MHI

Accuracy: 98%

Improvements: Significant boost in performance for actions like wave2 and side.

Key Enhancements: F1-score for wave2 improved from 0.94 to 0.99. Confusion reduced by over 80%.

Analysis: Optical flow helps track fast, subtle movements, improving accuracy.

Baseline vs Improvements (Continued)

Temporal Pyramid MHI
Accuracy: 97%
Key Observations: No significant improvement from baseline.
Analysis: Temporal segmentation may not effectively capture meaningful motion differences.

Dynamic MHI (DMHI)
Accuracy: 81%
Performance Drop: Severe misclassification, particularly with wave2 (all misclassified as bend).
Analysis: Likely issues with MHI construction, leading to noisy or distorted temporal dynamics.

Combined Optical Flow-Weighted MHI + Temporal Pyramid
Accuracy: 98%
Macro F1-score: 0.98
Best Performance: Balanced improvements across fast and complex actions.
Highlights: Improved stability and consistency, particularly for actions like "jump," with an F1 score increase from 0.95 to 1.00.
Key Advantage: Combines the strengths of optical flow and temporal structure to offer the best overall performance.

Key Takeaway:
Each enhancement adds more precision in motion tracking, leading to better action recognition, especially for dynamic and subtle movements.
The combination of optical flow and temporal pyramids yields the highest accuracy and the most robust system.

Bringing it all together

Our innovations were integrated at different stages to enhance action recognition beyond the baseline methods:

- Silhouette Masks: Isolating motion from the background for cleaner feature extraction and improved accuracy.
- MHI + Optical Flow: Capturing motion history and enhancing it with optical flow for better tracking of dynamic actions
- Hu Moments: Characterizing motion shapes, improving recognition of complex actions like walking or running.
- Projection Profiles: Summarizing motion distribution, adding spatial insight into how actions unfold.

How It All Comes Together:

- Feature Extraction: Innovations like silhouette masks, Hu Moments, and optical flow enhance motion representation.
- Model Training: Multiple variations (MHI + Flow-weighted, silhouette masks, Hu Moments) were tested to refine classification accuracy.
- Final System: The MHI + Flow-weighted method provided the best performance, shaped by insights from all innovations.

Conclusion

- We developed an enhanced action recognition system using MHI, silhouette masks, and advanced motion features.
- Innovations like silhouette masks, Hu Moments, and optical flow significantly improved motion representation and action classification.
- MHI + Flow-weighted delivered the highest accuracy, showing the power of combining motion history with optical flow for precise recognition.

Real-World Impact

- Enhances action recognition in fields like sports analytics, surveillance, and human-computer interaction.
- Offers a lightweight, efficient alternative to deep learning models for real-time applications.

Future Work

- **Handling Complex Backgrounds**: Improving robustness against cluttered environments and dynamic backgrounds.
- **Expanding Datasets**: Testing on more extensive and diverse datasets for better generalization and real-world performance.

Thank You!