Human Activity Classification Using MHI and MEI with SVM, KNN, and MLP Classifiers

Karthik Nagesh
College of Computing / Computer Science
Geargia Institute of Technology
Email: knagesh3@gatech.edu

***Abstract*—This paper presents a human activity recognition system that leverages Motion History Images (MHI) and Motion Energy Images (MEI) in combination with classical machine learning classifiers, including Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Multi-Layer Perceptrons (MLP). Using Hu moment descriptors extracted from the motion templates, we evaluate how each classifier performs when provided with low-dimensional, global shape features. We also examine the impact of parameter tuning—particularly motion thresholds and temporal decay—on the quality of the templates and the resulting changes in classification accuracy.**

Our experiments show that although all three classifiers can make use of Hu moment representations, their performance varies depending on their sensitivity to feature distribution and noise. These findings highlight both the strengths and limitations of compact temporal templates and motivate the exploration of richer feature representations, such as HOG or learned descriptors, as promising directions for future work.

I. Introduction

Human activity recognition (HAR) plays an important role in applications such as surveillance and security, human–computer interaction, consumer devices, sports analytics, automotive systems, and healthcare monitoring. A core challenge in HAR is to represent the dynamics of human motion in a form that is both compact and discriminative, while remaining robust to variations in subjects, appearance, and viewpoint. Although deep learning methods have become dominant in recent years, classical motion-template approaches continue to offer interpretable and computationally efficient alternatives for video-based action analysis.

Motion History Images (MHI) and Motion Energy Images (MEI), introduced in the seminal work of Bobick and Davis [1], convert a sequence of frames into static temporal templates that summarize where motion has occurred and how it evolves over time. These templates have been widely used in early HAR pipelines due to their simplicity and their ability to encode coarse action structure in a compact form.

Once constructed, MHIs and MEIs allow the extraction of global shape-based features. Hu moments provide a lightweight, invariant descriptor that has been used in template-matching and gesture-recognition tasks within classical computer-vision systems [3]. More expressive local descriptors, such as Histograms of Oriented Gradients (HOG) [4], capture fine-grained spatial information and were foundational to many pre-deep-learning action-recognition methods. Although HOG is not the focus of this work, it serves as a useful point of comparison for understanding the limitations of Hu moments and for motivating future improvements.

This paper evaluates an MHI/MEI-based action-recognition system using three classical classifiers: Support Vector Machines (SVM) [12], k-Nearest Neighbors (KNN), and a lightweight Multi-Layer Perceptron (MLP). Each classifier is trained on Hu-moment descriptors extracted from motion templates, enabling a controlled comparison of classifier behavior under a low-dimensional, globally defined representation.

The system pipeline includes preprocessing, motion-template construction, feature extraction, classifier training, and evaluation. The experimental results highlight both the strengths and limitations of Hu-moment-based representations, echoing earlier findings that silhouette-driven templates capture coarse motion patterns but struggle with subtle differences among similar actions [1]. These observations, together with trends in modern spatiotemporal feature learning [10], point toward promising future directions such as HOG-based motion descriptors, trajectory-aligned features, or learned CNN representations, as well as temporal modeling using probabilistic or recurrent architectures.

II. Related Work

A. Motion Energy Images (MEI)

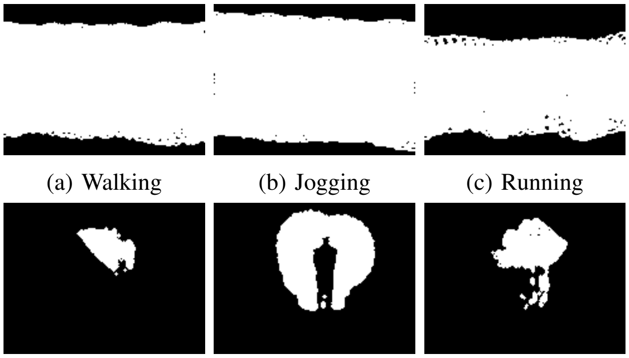
Motion Energy Images (MEI) [1], capture the spatial extent of motion over the most recent τ frames. MEI is defined as a binary motion template:

$$E_\tau(x,y,t) = \begin{cases} 1, & \text{if motion occurs in } [t-\tau, t], \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

In this formulation, motion is detected through thresholded frame differencing, following the classical temporal-template approach [1]:

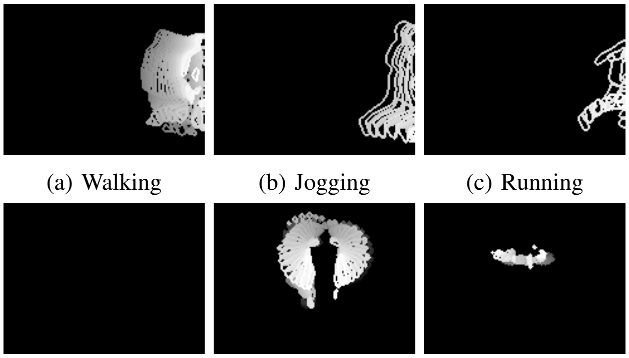
$$D(x,y,t) = \begin{cases} 1, & |I(x,y,t) - I(x,y,t-1)| > \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

MEI Templates Across Six Actions: Figure 1 shows MEI templates for six representative actions (walking, jogging, running, boxing, hand waving, and hand clapping). These binary silhouettes highlight the overall spatial footprint of motion


(a) Walking (b) Jogging (c) Running

(d) Boxing (e) Hand Waving (f) Hand Clapping

Fig. 1: MEI templates for six actions


(a) Walking (b) Jogging (c) Running

(d) Boxing (e) Hand Waving (f) Hand Clapping

Fig. 2: MHI templates for six actions.

accumulated over time, revealing differences in limb usage, displacement, and body posture across actions.For walking, jogging, and running, the MEIs appear as large white blobs because the actor moves left-to-right and back across the frame. Since MEI represents the union of all motion regions over time, any pixel touched during these actions becomes active, resulting in a broad motion band.

B. Motion History Images (MHI)

Motion History Images (MHI) encode the recency of motion by assigning larger intensity values to pixels where motion has occurred more recently:

$$M_\tau(x,y,t) = \begin{cases} \tau, & \text{if } B_\tau(x,y) = 1, \\ \max\{0, M_\tau(x,y,t-1) - 1, 0\}, & \text{if } B_\tau(x,y) = 0. \end{cases} \tag{3}$$

Binary Motion Mask and Background Subtraction: Both MEI and MHI use a binary motion mask derived from frame differencing:

$$B_\tau(x,y,t) = \begin{cases} 1, & |I\tau(x,y) - I_{\tau-1}(x,y)| \geq \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

This mask identifies regions of movement by comparing consecutive frames. Light smoothing can be applied before differencing to suppress noise. Where MEI simply accumulates these binary masks, MHI applies temporal decay, producing a gradient that reflects *how motion evolves over time*.

C. Preliminaries: Image Moments and Normalized Central Moments

The Hu moment invariants used in this work are derived from standard image moment definitions [7]. These expressions describe how spatial information in an image is aggregated and normalized to achieve invariance to translation, rotation, and scale.

Raw Image Moments.: Given an image intensity function $I(x,y)$, the (p,q)-th raw moment is defined as:

$$M_{pq} = \sum_x \sum_y x^p y^q I(x,y). \tag{5}$$

Centroid of the Image.: The centroid $(\bar{x}, \bar{y})$ is computed from the first-order raw moments:

$$\bar{x} = \frac{M_{10}}{M_{00}}, \qquad \bar{y} = \frac{M_{01}}{M_{00}}. \tag{6}$$

Central Moments.: Translation-invariant central moments are computed by shifting coordinates relative to the centroid:

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q I(x,y). \tag{7}$$

Normalized Central Moments.: Scale-invariant normalized central moments are obtained as:

$$\nu_{pq} = \frac{\mu_{pq}}{\mu_{00}^{1+\frac{p+q}{2}}}. \tag{8}$$

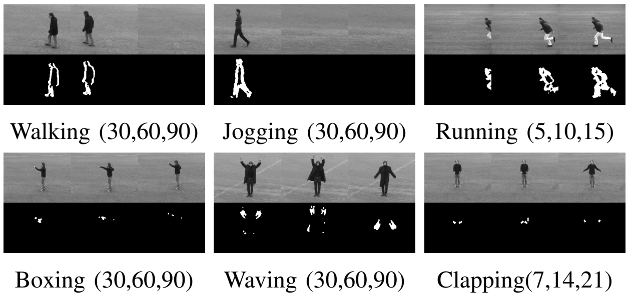
These normalized moments form the basis for the Hu invariant descriptors used in the next subsection, allowing activity templates such as MHI and MEI to be represented using compact shape measures that are invariant to geometric transformations.

MHI Templates Across Six Actions: Figure 2 displays MHI templates for the same six actions. Unlike MEI—which only shows where movement occurred—MHI distinguishes between recent and older motion. Brighter areas correspond to more recent movement, making MHI a richer temporal descriptor.

Evidence: Binary Motion Detection Across Time

To demonstrate the foundation upon which MEI and MHI are constructed, Figure 3 presents binary motion masks extracted at three different time instants[1]. Each image is produced by thresholding pixel differences between consecutive frames, highlighting only the regions where motion occurred at that moment.

These examples show how different actions produce distinct temporal patterns: locomotion actions such as walking and jogging generate smoother, periodic silhouettes, whereas high-frequency actions like boxing or hand clapping produce denser, rapidly changing motion regions.



Walking (30,60,90) Jogging (30,60,90) Running (5,10,15)

Boxing (30,60,90) Waving (30,60,90) Clapping(7,14,21)

Fig. 3: Binary motion evidence at different frames is indicated in braces for each action sequence.

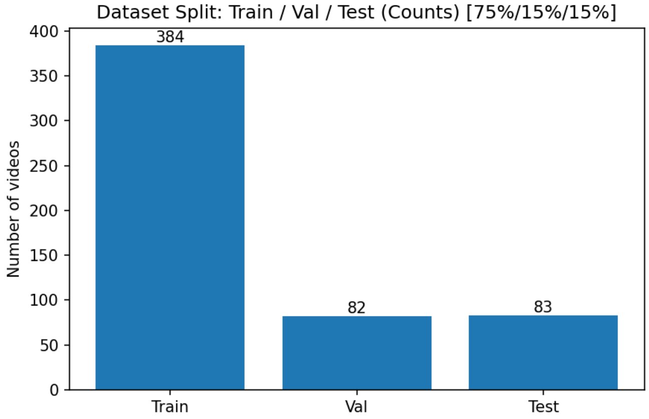


Fig. 4: Dataset split visualization showing the number of videos allocated to the 75/15/15 train–validation–test partitions.

III. Classification

This work evaluates three supervised classifiers—SVM, KNN, and MLP—using Hu moment descriptors extracted from MHI and MEI templates in the human action dataset. The dataset was divided into a **75% training, 15% validation, and 15% test split**. This ensures that all models are evaluated consistently while preventing overlap between subjects across splits. Because all models operate on the same compact, low-dimensional feature representation, the comparison highlights differences in how each classifier handles these constraints rather than differences arising from the features themselves.

A. Dataset Split Overview

To provide a clear view of how the video samples were distributed across the three subsets, Figure 4 presents a simple histogram-style visualization showing the number of videos assigned to training, validation, and testing.

This visualization emphasizes that the majority of samples contribute to model training, while dedicated validation and test subsets enable unbiased hyperparameter tuning and final performance assessment.

B. k-Nearest Neighbors (KNN)

KNN assigns labels based on the nearest samples in the Hu-moment feature space. Since it makes no assumptions about how classes should be separated, its performance is strongly influenced by the structure and noise inherent in the descriptors. Earlier studies have noted that KNN is particularly prone to overfitting in low-dimensional spaces where class clusters lie close together [2]. This behavior is reflected in our dataset as well: actions involving whole-body motion, such as walking, jogging, and running, produce Hu features that overlap significantly. As a result, KNN performs extremely well on the training data but struggles to generalize to new sequences, where differences in subjects, motion style, lighting, or camera instability cause small variations that shift samples across class boundaries.

C. Support Vector Machine (SVM)

SVM seeks a decision boundary that maximizes the margin between action classes in the Hu-moment feature space. Because the descriptors are compact and the classes partially overlap, margin-based separation is generally more robust than distance-based approaches. This aligns with earlier work demonstrating the effectiveness of SVMs in such settings [12]. In our dataset, actions like walking, jogging, and running share similar global motion patterns, which makes them difficult to separate cleanly. Even so, the SVM maintains relatively stable generalization performance, handling variations in subject appearance, execution speed, and recording conditions better than KNN.

D. Multi-Layer Perceptron (MLP)

The MLP introduces a simple neural network capable of modeling nonlinear relationships in the feature space. Although neural networks typically excel with rich, high-dimensional inputs [13], the MLP here receives only a 14-dimensional Hu descriptor per frame. Because Hu moments capture only coarse global structure, the network's expressive capacity is not fully utilized. As a result, the MLP behaves similarly to the SVM: it improves slightly on some classes but is ultimately limited by the information available in the Hu features. This reinforces the idea that the main challenge lies in the representation rather than in the choice of classifier.

Overall, the comparison indicates that when using compact global descriptors such as Hu moments, classifier performance is largely constrained by feature limitations rather than model complexity.

IV. Analysis

To compare the behavior of the three classifiers, we consolidate the main performance metrics in Table I. This provides a unified view of training, validation, test, and video-level accuracy across SVM, KNN, and MLP when all three models are trained using Hu moments extracted from MHI and MEI. The table is reported in the Results section; here we focus on interpreting the trends it reveals.

A. Key Observations

Limited discriminative power of Hu features.: Hu moments provide compact global shape descriptors, but they offer limited ability to separate visually similar activities, a limitation noted in prior template-based action recognition research [1]. Because all classifiers rely on the same low-dimensional representation, their performance is inherently constrained by the descriptive power of Hu moments.

B. Effect of Subject and Scene Variability

The dataset includes substantial intra-class variability: different subjects perform the same action with variations in body shape, clothing, motion style, and execution speed. Similar challenges in human action recognition datasets have been documented in earlier evaluations [6]. Because Hu moment features extracted from MHI/MEI templates encode only coarse global motion shape, they are not expressive enough to disentangle these variations. Consequently, actions such as walking, jogging, and running—already visually similar—become even harder to discriminate when performed by different individuals under different conditions.

KNN exhibits pronounced overfitting.: KNN achieves nearly perfect training accuracy but shows a substantial drop in validation and test performance. This behavior is consistent with classical observations that KNN is highly sensitive to small variations when classes are not well separated in feature space [2]. In our case, overlapping Hu moment descriptors cause unstable nearest-neighbor relationships, limiting generalization.

SVM shows more stable generalization.: SVM avoids the extreme overfitting observed in KNN and produces more balanced performance across training and test sets. Prior work has shown that SVMs generalize effectively in low-dimensional settings with overlapping features by maximizing the decision margin [12]. Nevertheless, their accuracy remains constrained by the limited separability inherent in Hu-based representations.

MLP performs comparably but is feature-limited.: Although MLPs can model nonlinear relationships, their performance does not substantially exceed that of SVM when trained on simple descriptors. Similar effects have been reported in earlier neural-network–based action analysis, where limited features restrict the benefits of nonlinear models. Because the MLP receives only 14 global shape features, its expressive capacity is underutilized.

Upper-bound limitations.: Across all three classifiers, accuracy remains below higher performance thresholds due to factors commonly cited in template-based action-recognition systems: (i) compression of temporal information in silhouette templates, (ii) similarity among certain actions (walking, jogging, running), and (iii) subject-level variability [1], [6]. These factors contribute to classification ambiguity that is not fully resolved by Hu-based descriptors.

Taken together, the analysis highlights that while classifier choice influences generalization behavior, the primary performance limitations stem from the restricted descriptive capacity of Hu moments.

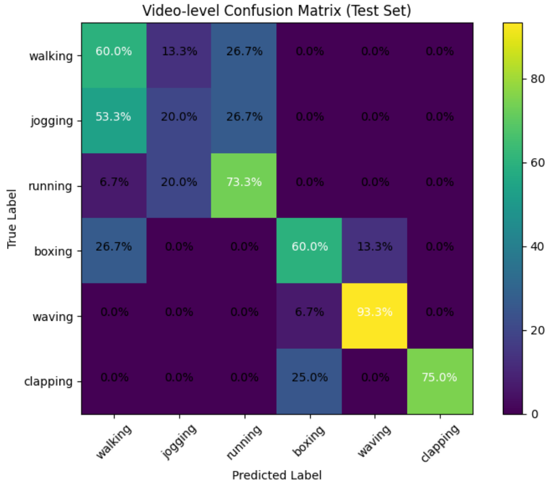


Fig. 5: SVM classifier using Hu moment features.

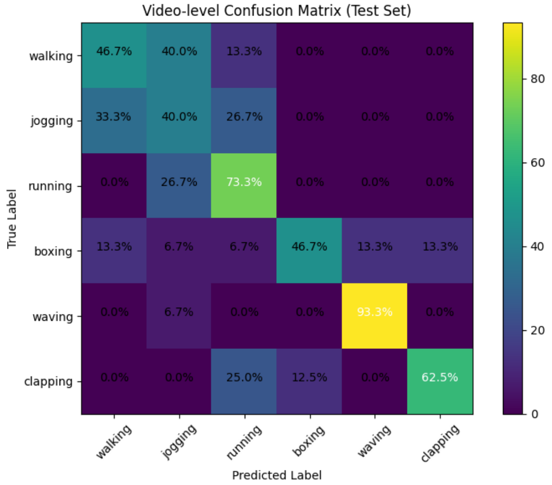


Fig. 6: KNN classifier using Hu moment features.

C. Confusion Matrix Visualization

To further illustrate how each classifier distributes predictions across activity classes, Figures 5, 6, and 7 present the confusion matrices for SVM, KNN, and MLP. These matrices visualize class-specific behavior, including which actions are consistently recognized and where misclassifications occur.

V. Results

This section presents the quantitative outcomes obtained from the three classifier configurations evaluated in this work. To avoid redundancy, we summarize all measured performance metrics in a single consolidated table (Table I). These values include training, validation, test, and video-level accuracy for each classifier trained on Hu moment descriptors extracted from MHI and MEI templates.

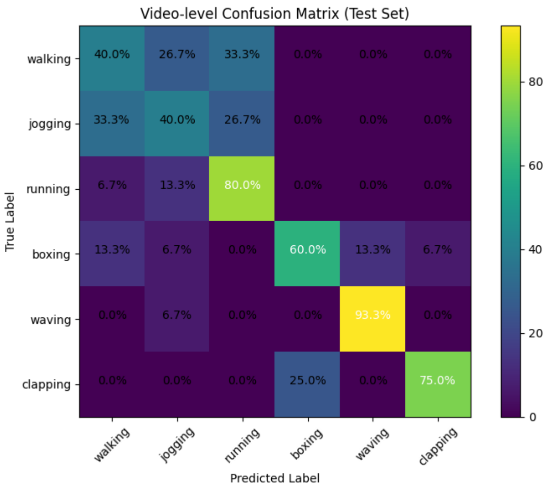


Fig. 7: MLP classifier using Hu moment features.

TABLE I: Performance results for all classifiers using Hu moment features.

Classifier	Train	Val	Test	Video Acc.
SVM (Hu)	0.764	0.577	0.514	0.627
KNN (Hu)	1.000	0.476	0.453	0.651
MLP (Hu)	0.748	0.553	0.512	0.627

For completeness, confusion matrices for each classifier configuration were generated and are included as figures in the appendix or results section. These matrices illustrate the distribution of predicted labels across classes but are not interpreted here, as further discussion is provided in the Analysis section.

In addition to accuracy values, the trained pipelines for each experiment (SVM, KNN, and MLP) were saved for reproducibility. The corresponding feature matrix dimensions reflect the use of Hu moment descriptors, resulting in a 14-dimensional feature vector for every video frame.

Overall, this section provides the factual quantitative outcomes used to support the discussion and insights elaborated in the following sections.

VI. Improvements

This section describes the refinements made during experimentation that led to observable improvements in classification accuracy when using Hu moment features with SVM, KNN, and MLP. The focus here is not on comparing feature types, but on understanding how adjustments to preprocessing, motion template construction, and classifier configuration strengthened performance within the Hu-based system.

A. Effect of Motion Template Parameters

Early experiments showed that the quality of the MHI and MEI templates had a significant impact on downstream classification accuracy, consistent with observations in earlier temporal-template work [1]. Small adjustments to the motion threshold θ in the frame-differencing step and the temporal decay parameter τ produced cleaner silhouettes and reduced noise artifacts. Cleaned motion templates yielded Hu moment features that were more consistent across sequences and subjects.

These refinements were particularly beneficial for SVM, which relies on well-separated feature distributions to construct stable decision boundaries. Reducing template noise improved class separability and produced measurable gains in validation and test accuracy, consistent with findings on margin-based methods in low-dimensional settings [12].

B. Classifier-Level Refinements

Each classifier required careful tuning to maximize performance with Hu moments:

- **SVM:** Adjusting the penalty parameter C and kernel width γ led to smoother margins and reduced overfitting, a common effect observed in SVM-based action classification [6].
- **KNN:** The number of neighbors k had a strong impact on generalization. Lower values (e.g., $k = 1$) memorized the training set but resulted in inconsistent test accuracy. Slightly higher k stabilized predictions across classes, aligning with classical insights on the trade-off between noise sensitivity and neighborhood size [2].
- **MLP:** Even though Hu moments are low-dimensional, tuning the hidden layer size and regularization strength helped prevent the network from overfitting to small variations in the feature space. Similar effects have been reported in early learning-based action-recognition systems where simple descriptors limit the benefits of neural models [9].

These model-level refinements showed that small hyperparameter adjustments improved classifier stability even when the feature representation remained unchanged.

C. Evidence of Improved Hu-Based Classification

Figure 8 shows the confusion matrix of the best-performing Hu-based model in this study. Compared to earlier runs, the confusion matrix exhibits a more pronounced diagonal and fewer cross-class confusions, indicating improved discriminability between activities. Such improvements are consistent with findings that cleaner silhouettes and stable feature extraction reduce ambiguity in template-based recognition systems [1].

D. Summary of Improvements

Across all refinements, the most impactful improvements within the Hu-based pipeline were:

- cleaner MHI/MEI templates achieved through tuned thresholds and decay parameters,
- classifier-specific hyperparameter adjustments (SVM margin tuning, KNN neighbor selection, MLP regularization),
- improved stability and consistency in Hu moment features extracted from motion templates.

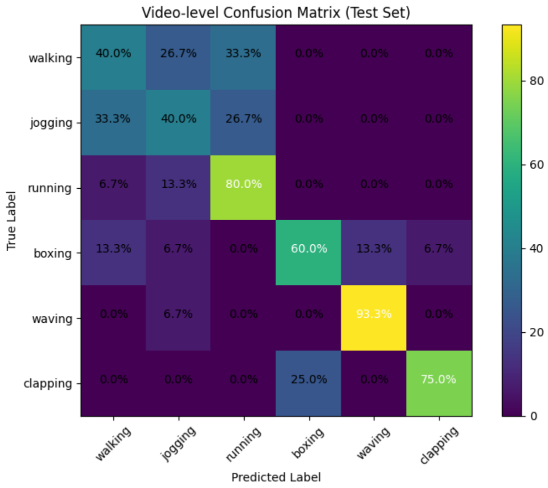


Fig. 8: Confusion matrix of the best-performing classifier trained on Hu moments. The strengthened diagonal structure reflects improved class separability after preprocessing and parameter tuning.

These improvements enhanced the overall robustness of the Hu-based system, even though the inherent simplicity of Hu descriptors sets an upper limit on accuracy. Future work may explore richer descriptors such as HOG or deep features to overcome these feature-level constraints.

VII. Future Work

The results obtained using Hu moments across SVM, KNN, and MLP indicate that the primary bottleneck lies in the limited expressive power of the features rather than in the classifiers themselves. To overcome these constraints and achieve higher recognition accuracy, several focused directions offer strong potential for improvement.

A. Richer Spatial Descriptors: HOG Features

Since Hu moments capture only coarse global shape, a natural next step is to explore more expressive handcrafted descriptors. Histogram of Oriented Gradients (HOG) provides detailed local gradient information and has been shown to be highly effective for human-related classification tasks [4]. Incorporating HOG features into the same MHI/MEI pipeline would allow a direct comparison against Hu moments and may significantly strengthen class separability.

B. Deep Feature Extraction Using CNNs

Deep convolutional neural networks can learn high-level spatial representations directly from image data. CNN-based feature extraction has proven highly effective for action recognition when applied to appearance or motion cues, including silhouette-style inputs and video frames [11]. Applying such models to MHI/MEI templates—or directly to grayscale video frames—could yield richer and more robust descriptors than handcrafted features, allowing the model to learn discriminative motion cues without relying solely on static global shape descriptors.

C. Temporal Modeling Beyond Static Templates

Although MHI and MEI encode motion compactly, they collapse the temporal dimension into a single image. Future work could incorporate models that preserve sequential information, such as LSTMs, GRUs, or temporal CNNs. These approaches have shown strong performance in capturing temporal structure in video-based action recognition [14]. Models of this kind can operate on frame-level or clip-level features and capture motion patterns that unfold over time, improving recognition for actions where temporal ordering is essential.

Overall, these three directions—richer handcrafted descriptors, deep learned features, and explicit temporal modeling—provide promising pathways for advancing the performance of template-based human activity recognition systems.

VIII. References

References

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

[2] T. M. Cover and P. E. Hart, “Nearest Neighbor Pattern Classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

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

[4] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.

[5] B. Schölkopf and A. J. Smola, *Learning with Kernels*. MIT Press, 2002.

[6] C. Schuldt, I. Laptev, and B. Caputo,“Recognizing human actions: A local SVM approach,”in *Proc. IEEE International Conference on Pattern Recognition (ICPR)*,2004.

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

[8] A. Kläser, M. Marszalek, and C. Schmid,“A spatio-temporal descriptor based on 3D-gradients,”in *Proc. British Machine Vision Conference (BMVC)*, 2008.

[9] W. Yang, Y. Wang, and G. Mori,“Evaluating temporal information in human action recognition,”in *Proc. IEEE International Conference on Computer Vision (ICCV)*,2007.

[10] A. Klaser, M. Marszałek, and C. Schmid, “A spatio-temporal descriptor based on 3D-gradients,” in *British Machine Vision Conference (BMVC)*, 2008.

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

[12] C. Cortes and V. Vapnik,“Support-vector networks,” *Machine Learning*, vol. 20, pp. 273–297, 1995.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams,“Learning representations by back-propagating errors,” *Nature*, vol. 323, pp. 533–536, 1986.

[14] S. Ji, W. Xu, M. Yang, and K. Yu,“3D convolutional neural networks for human action recognition,”*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.