Combining Shape of Trajectories with MHI and their Directional Derivative-Based Description for Human Activity Recognition

Original Scientific Paper

Siddharth Bhorge*
Vishwakarma Institute of Technology,
Department of Electronics and Telecommunication
Pune, Maharashtra, India
siddharth.bhorge@vit.edu

Medha Wyawahare
Vishwakarma Institute of Technology,
Department of Electronics and Telecommunication
Pune, Maharashtra, India
medha.wyawahare@vit.edu

Vijay Mane
Vishwakarma Institute of Technology,
Department of Electronics and Telecommunication
Pune, Maharashtra, India
vijay.mane@vit.edu

Milind Kamble
Vishwakarma Institute of Technology,
Department of Electronics and Telecommunication
Pune, Maharashtra, India
milind.kamble@vit.edu

Milind Rane
Vishwakarma Institute of Technology,
Department of Electronics and Telecommunication
Pune, Maharashtra, India
milind.rane@vit.edu

*Corresponding author

Abstract *– This research introduces a unified framework for human activity recognition that integrates global temporal characteristics, local spatial information, and trajectory shape cues. Trajectory shapes are extracted by tracking key points using a Motion History Image (MHI) as a mask, eliminating the need for unreliable key-point and trajectory tracking. The selected key points from both the intensity image (local spatial information) and the MHI (global temporal information) are represented using the Histogram of Directional Derivative (HODD) descriptor, which effectively captures their visual and structural attributes. The combined feature representation is encoded through a Bag-of-Visual-Words (BoVW) model, and classification is performed using a multiclass Support Vector Machine (SVM). Extensive experiments on four benchmark datasets—URADL, KTH, Weizmann, and UCF101—yield accuracies of 95.4%, 95.83%, 100%, and 89%, respectively, demonstrating robustness to illumination changes, occlusion, and background clutter, and outperforming several state-of-the-art methods. Overall, the proposed framework offers a computationally efficient and highly discriminative solution for human activity recognition by effectively fusing trajectory shape, spatial, and temporal descriptors.*

Keywords: *Histogram of directional derivative, human activity recognition, MHI, and shape of trajectories*

Received: June 20, 2025; Received in revised form: December 7, 2025; Accepted: December 7, 2025

1. INTRODUCTION

In the field of computer vision, Human Activity Recognition (HAR) is highly significant due to its diverse applications in video analysis [1], automated surveillance [2], human–computer interaction [3], and elderly care. Multiple studies highlight the importance of providing effective solutions for senior individuals who live independently and alone. However, HAR is still considered a challenging task within the computer vision community due to the heterogeneous nature of video sequences. Furthermore, representing videos captured in unconstrained environments poses several challenges, including occlusion, viewpoint variation, and background clutter.

A wide range of methodologies—from statistical models to deep learning networks—have been developed for HAR. Generally, HAR techniques can be categorized into two major groups: global feature–based approaches and local feature–based approaches. Local feature–based

techniques operate on video sequences using spatio-temporal interest point detectors and descriptors, whereas global feature–based techniques utilize appearance, direct motion patterns, silhouette information, and shape cues. Among these, local spatio-temporal feature–based methods [4], [5], [6] have achieved remarkable progress, often requiring comparatively less complex preprocessing. These features are especially effective because they encode rich motion information from the video.

In recent years, multiple-feature fusion techniques [7], [8] have gained attention, as they often outperform single-feature approaches. Despite the promising advancements, more robust techniques are still needed to address (i) unreliable tracking of key points caused by object motion unrelated to the primary action, and (ii) the difficulty in representing complex activities under noise and varying illumination conditions.

To address these issues, we propose a novel activity recognition framework with the following key contributions:

- To mitigate unreliable key-point tracking and trajectory instability, we employ Motion History Images (MHI) as a masking technique.
- To effectively represent complex activities, we combine trajectory-based local motion information with spatial key-point intensity features and temporal MHI features using the HODD descriptor.

An outline of the paper is as follows: Section 2 presents a review of existing descriptors. Section 3 describes the proposed feature extraction method and the integration of heterogeneous feature descriptors. Section 4 provides experimental results on the URADL activities-of-daily-life dataset [9], the KTH dataset [10], and the Weizmann dataset [11]. Section 5 concludes the paper.

2. LITERATURE REVIEW

Over the past decade, numerous computer vision techniques and strategies have emerged within the field of Human Activity Recognition (HAR). These methods rely on discriminative visual features extracted from action sequences and are broadly categorized into global and local representations. Global feature-based techniques use shape, appearance, silhouette cues, and direct motion information. Bobick and Davis [12, 13] introduced Motion History Images (MHI) for representing global spatio-temporal motion, where templates are constructed by aggregating the temporal history of motion at individual pixels to capture overall motion dynamics. Du-Ming Tsai *et* al. [14] extended this idea by constructing MHI using optical flow magnitude at each pixel. Islam S. *et* al. [15] utilized silhouette shape as a global descriptor, applying optical flow along the silhouette boundary to encode motion. Global descriptors are computationally efficient and simple to implement because they avoid explicit keypoint tracking; however, they are highly sensitive to lighting variations, background clutter, occlusion, and viewpoint changes, which can distort silhouette and motion cues.

To overcome these limitations, a second major class of action representation focuses on local features [16-18]. These approaches analyze localized regions to capture motion, appearance, and spatio-temporal characteristics that are often missed by holistic descriptors. Local techniques frequently outperform global ones, particularly in cluttered or dynamic environments. Trajectory-based methods [19-24] have gained considerable attention, as they track key points across successive frames and extract long-term motion patterns that provide powerful cues for action differentiation. Despite their advantages, local and trajectory-based approaches may capture insufficient structural context, are computationally demanding, and can perform poorly when motion is weak or uniform. To address these drawbacks, hybrid techniques [25, 26] combine local representations with global structural information.

Many hybrid models demonstrate the effectiveness of integrating complementary cues. Ahmad M. and S. W. Lee [27] combined motion and shape features using invariant moments and global optical flow. Tian *et* al. [28] fused local intensity-based interest points with global MHI features, enabling effective action recognition in crowded scenes. Zhao D. *et* al. [29] represented structural and appearance information using HOG3D to describe spatio-temporal interest points. Yu *et* al. [30] combined keypoint trajectories with local descriptors such as HOG and MBH. Luvizon D. C. *et* al. [31] merged spatial and local temporal features using a metric learning framework. These hybrid and fusion-based frameworks produce more robust and discriminative representations by leveraging the strengths of both global structure and local motion cues, although they often require large datasets, careful feature weighting, and may be susceptible to redundancy or overfitting.

In recent years, significant advances in HAR have been driven by deep learning, transformer-based video models, multimodal fusion, and vision–language foundation architectures. Tong *et* al. [32] introduced a self-supervised training strategy for video data using heavy spatio-temporal masking on Vision Transformers (ViT). This approach is data-efficient and transferable to small datasets but requires substantial computational resources and is sensitive to masking ratios. Garg *et* al. [33] proposed a hybrid CNN–LSTM framework that effectively integrates spatial and temporal cues, achieving strong performance on KTH and UCF datasets but demonstrating limited scalability to complex real-world scenes. Morshed *et* al. [34] presented a structured taxonomy classifying HAR methods into handcrafted, hybrid, and deep learning approaches, highlighting emerging multimodal trends but offering limited critique of transformer-based models. Sánchez-Caballero *et* al. [35] developed a ConvLSTM-based model optimized for depth-only HAR, providing low-latency and privacy-preserving inference suitable for embedded

systems, though dependent on depth sensor quality and less effective for appearance-based tasks. Hu *et* al. [36] proposed an attention-based fusion network combining RGB, skeleton, and optical flow data, yielding improved robustness but incurring high runtime complexity and requiring multimodal data availability. Zhang, Li, and Xu [37] adapted CLIP and other foundation models for HAR to enable zero-shot and few-shot learning, improving generalization and label efficiency at the cost of significant computational overhead and challenges in modeling fine-grained temporal details.

Recent research further highlights the value of multimodal fusion, hierarchical modeling, and interpretable handcrafted features. Kamble and Bichkar [38] presented a hierarchical framework that models body-part interactions to improve semantic understanding of complex activities. Their architecture uses region-level motion cues to infer full-body actions, offering enhanced interpretability and strong performance under occlusion but requiring accurate pose estimation and segmentation. Compared with their method, our proposed approach captures integrated spatio-temporal information from trajectories and MHI descriptors without explicit segmentation, providing better generalization across viewpoints.

S. Abraham and R. K. James [39] investigated the use of handcrafted spatio-temporal features combined with attention-based RNN models, showing that handcrafted descriptors can complement deep learning architectures to provide interpretability and competitive accuracy. Their method, however, depends on high-quality feature engineering and struggles with long-term temporal dependencies—challenges that our approach addresses by integrating trajectory-based motion cues with temporal MHI structure.

Finally, D. S. Korti and Z. Slimane [40] proposed an HAR framework using micro-Doppler radar signatures, employing feature concatenation and augmentation for robust recognition in noisy and privacy-sensitive environments. Although radar-based, their findings reinforce the importance of multi-feature fusion and robust representation learning, aligning with our motivation to combine complementary spatial and temporal cues for improved action recognition.

3. METHODOLOGY

The suggested action recognition framework is shown in Fig. 1.

It. enhances Human Activity Recognition by effectively combining trajectory-based features with both local (spatial) and global (temporal) features. The framework consists of three main components—feature extraction, feature description, and action recognition.

A technique presented by Tian Y. *et* al. [29] is used in the feature extraction procedure to efficiently separate spatial and temporal data. Temporal information is generated through the creation of an MHI [12, 13], while spatial information is extracted from 2D Harris corners in the initial images. In the feature descriptor stage, we propose integrating trajectory-based features with local features (2D Harris corner points in intensity images) and global features (2D Harris corner points in MHI images) to form a unified final feature vector. The global and local features are characterized using the newly adopted HODD descriptor [41] in the MHI and intensity images, respectively. To compute trajectories, 2D Harris corner points are tracked as key points using the traditional Lucas–Kanade–Tomasi (KLT) tracker [42].

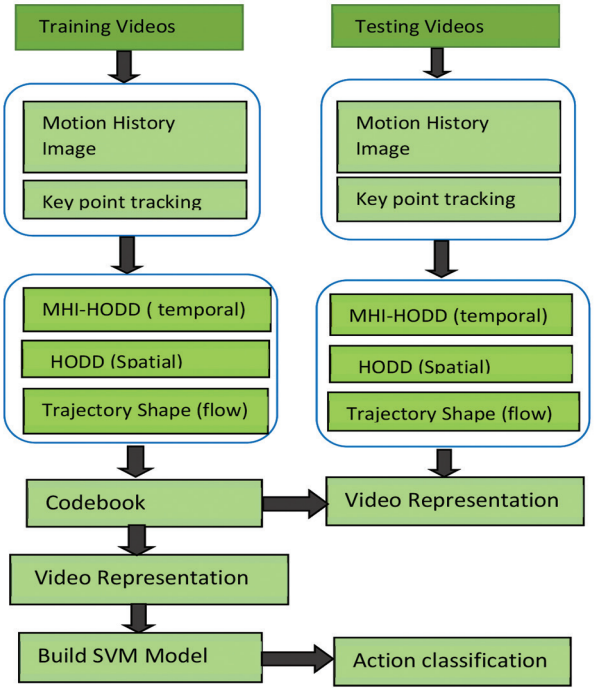


Fig. 1. Proposed system for human action Recognition

To address the problem of inconsistent tracking of key points, the MHI is utilized as a masking mechanism, ensuring that only action-related motions contribute to the trajectory computation. These trajectories are then characterized based on their distinct motion-shape attributes. The resulting feature descriptor is created by combining three components:

(i) HODD descriptors of key points in the intensity image, (ii) HODD descriptors of key points extracted from the MHI ,and (iii) trajectory-shape features derived from the tracked key points.

Subsequently, k-means clustering is applied to all extracted descriptors to construct a Bag-of-Visual-Words (BoVW) model for the training videos. Using Euclidean distance as the similarity metric, the cluster centers (visual words) represent each feature descriptor. Consequently, each training video sequence is encoded as a BoVW histogram. The testing phase follows the same procedure as the training phase, where each testing video is assigned a BoVW representation using the precomputed visual vocabulary.

In the final stage, a Support Vector Machine (SVM) classifier is employed to categorize the testing video sequences. The subsequent subsections provide a detailed description of the complete HAR framework.

3.1. FEATURE EXTRACTION

3.1.1. Motion history image (MHI)

The MHI functions as a compact motion template for static 2D images that is derived from spatio-temporal (3D) information in an image sequence. The MHI compresses temporal motion patterns into a single image, where the intensity of each pixel corresponds to the recency and frequency of motion. It is generated by computing consecutive frame differences, which are accumulated to form temporal layers merged into one static intensity image.

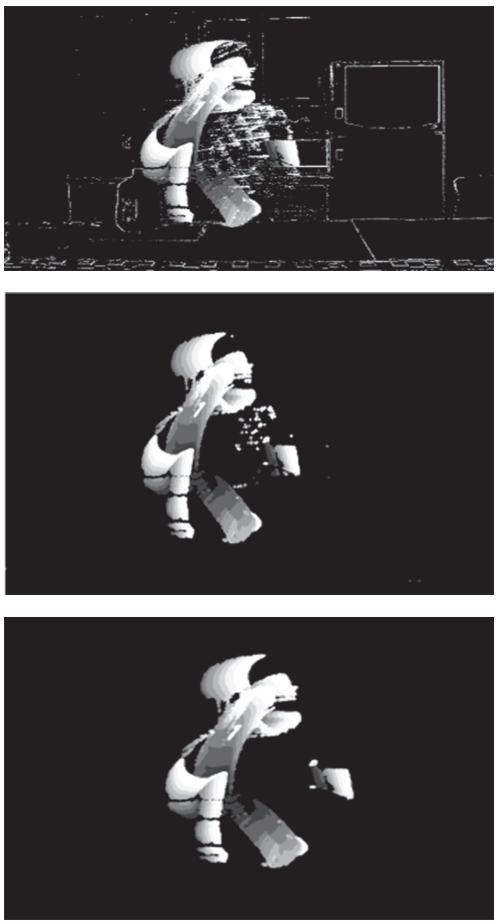
MHI is constructed using the following definition given in [12]

$$MHI_{\tau} = \begin{cases} \tau, if\ I(i,j,t) = 1 \\ max(0, MHI_{\tau}(i,j,t-1) - 1), otherwise \end{cases} \quad (1)$$

In this formulation, the binary image $I(i, j, t)$ is computed as the frame-difference map between successive images, and the parameter τ represents the temporal duration of motion accumulation. MHI is known to be robust against noise, partial occlusion, missing body parts, and shadows [43]. Since it encodes the entire temporal motion history into a single 2D matrix, the representation is computationally efficient and well-suited for dynamic motion characterization. The construction process of the MHI is illustrated in Fig. 2. Fig. 2(a) represent a sample frame depicting the activity 'answering a phone call'. The resulting MHIs, however, may contain noise, making preprocessing and noise-removal steps essential before further analysis.

The construction of the MHI is shown in Fig. 2. Fig. 2(a) illustrates a representative frame of the action "drinking water," while Fig. 2(b) displays the corresponding MHI. The presence of noise in the generated MHIs highlights the need for preprocessing to remove unwanted artifacts. First, an opening operation is applied to eliminate background noise caused by variations in illumination and lighting conditions. Next, blob analysis is performed to determine the size of each connected component in the MHI. Blobs with an area of approximately 10×10 (100) pixels are removed to suppress insignificant motion regions. The final, noise-reduced MHIs obtained after these preprocessing steps are shown in Fig. 2(d)

(a)



(b) (c) (d)

Fig. 2. Illustration of construction of MHI (**a**) representative frame of drinking water (**b**) MHI of drinking water. (**c**) MHI after opening operation. (**d**) MHI after blob analysis

3.1.2. Detection of Key Points

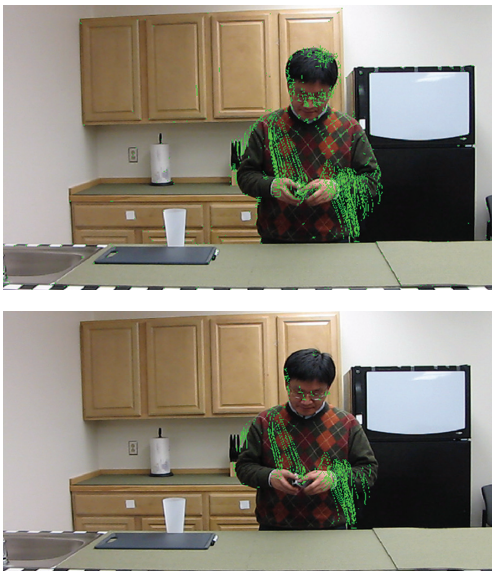
Over the past decade, several spatio-temporal interest point detectors have emerged in computer vision, demonstrating strong performance in action recognition tasks [16]. We utilized the well-known 2D Harris corner detector [44] to extract interest points from the video sequence. Due to its invariance to illumination changes, rotation, and moderate scale variations, it becomes more suitable for action recognition. The local auto-correlation of the image function $I(x, y)$ is calculated to determine the interest points. In this approach, a little shift in either direction of the image function $IM(i, j)$is analysed.

It is defined as follows.

$$C(u,v) \rightleftarrows = \sum_{i,j} M(i,j) \\ [IM(i+u, j+u) - IM(i,j)]^2 \quad (2)$$

Following the detection of key points, a standard KLT tracker is employed to track them across consecutive video frames. To address issues with unreliable key points and noisy or unstable trajectories, the MHI is used as a masking mechanism during tracking. In our implementation, key points are tracked over a range of 15 to 30 frames.

Fig. 3(a) shows the trajectories of tracked key points without MHI as mask for the activity "answer a phone call." Fig. 3(b) displays the trajectories of key points with the MHI mask applied, and it shows that the unreliable trajectories are effectively removed.



(a)

(b)

Fig. 3. Example of tracked key points with and without MHI a mask (**a**) tracked key points without MHI (**b**) tracked key points with MHI mask

3.2. FEATURE DESCRIPTOR

3.2.1. HODD and MHI-HODD feature descriptor

Within the literature, Histogram of Oriented Gradient (HOG) and Histogram of Optical Flow (HOF) feature descriptors are extensively utilized in human detection and action recognition [45-47]. HOG constructs a histogram of gradient vectors by quantizing their orientations, effectively capturing structural details along edges. Motivated by the limitations of HOG and HOF in capturing richer directional information, we implemented HODD, which extracts not only appearance and structural information but also directional information along multiple orientations defined by a unit vector, in addition to the normal direction. In our proposed framework, HODD describes the local structure and appearance characteristics of both the MHI and the intensity image. The calculation of HODD is performed by assessing the directional derivative along a unit vector v, as presented in the following equation:

$$DD_v f(a) = \langle \nabla f(a), v \rangle \tag{3}$$

$$\bar{v} = \bar{a}_x \cos\,\theta + \bar{a}_y \sin\,\theta - \pi < \theta < \pi \tag{4}$$

Each key point in the MHI and the intensity image is characterized by a neighbourhood window of size (W_x, W_y)centered at the key point. This window is further divided into non-overlapping sub-windows of size (m_x, m_y). In the proposed method, we select only those tracked key points exhibiting recent motion (MHI intensity > threshold) for descriptor computation, ensuring the use of motion-relevant regions. The local appearance and motion information are then described using HODD at each selected key point in both the MHI (illustrated in Fig. 4) and the intensity image.

In our experiments, we set the window size to **32 × 32** and the sub-window size to **8 × 8**. We utilized **9** bins for the HODD computed on the intensity image and an additional 9 bins for the HODD computed on the MHI. The final feature vector was formed by concatenating and normalizing the histograms of all sub-windows, resulting in a compact yet discriminative representation.

3.2.2. Trajectory shape descriptor

To depict the shape of trajectories, we adopted a shape descriptor introduced by Wang *et* al. [47], which is specifically designed to capture fine-grained local motion variations along the trajectory path. In our approach, interest points were monitored over a timeframe ranging from 15 to 30 frames. Describing the trajectory's shape, with a length of L, involves representing it through a sequence of displacement vectors (ΔV_t,…,ΔV_{t+1}), where:

$$\Delta V_t = (V_{t+1} - V_t) = (i_{t+1} - i_t, j_{t+1} - j_t) \tag{5}$$

The length of the feature vector is therefore $2{\times}L$ for a trajectory consisting of L points. By integrating the normalized appearance and structural descriptors (HODD from the intensity image) with the temporal descriptor (MHI-HODD) and the motion descriptor (trajectory shapes), we construct a unified composite feature vector that captures spatial, temporal, and motion-specific information in a complementary manner.

3.3. ACTION CLASSIFICATION

In Sections 3.1. and 3.2., we illustrated the extraction and description of features through the use of MHI-HODD and the shape of trajectories, which collectively capture structural, appearance, and motion information. Action classification is performed using a BoVW model alongside a multiclass SVM [48] classifier. Both the MHI-HODD descriptors and the trajectory-shape descriptors are used to construct visual vocabularies, allowing each video sequence to be represented as a frequency histogram of feature occurrences. In the training videos, histograms are generated for all vocabulary words. These histograms are then used to train an SVM model capable of recognizing actions in the test videos.

We used a leave-one-out cross-validation (LOOCV) strategy for classification in our experiment. Within this scheme, samples from a single class are iteratively used as test data, while the remaining samples constitute the training set, and predictions are made using the trained model. Subsequently, for M classes, an $M{\times}M$ confusion matrix is created, comparing the input class samples with the predicted class labels and categorizing them as false positive (FP), false negative (FN), true positive (TP), and true negative (TN). The overall classi-

fication performance is quantified using the following expression for average accuracy:

$$\text{Average Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

4. RESULT AND DISCUSSION

4.1. RESULTS ON URADL DATASET

The URADL dataset contains 150 high-resolution indoor video clips representing ten Activities of Daily Living (ADL). Five individuals perform each activity three times, and all recordings were captured using a static RGB camera to support the evaluation of action-recognition methods. The activities include: Answer Phone (A1), Chop Banana (A2), Eat Snack (A3), Dial Phone (A4), Drink Water (A5), Peel Banana (A8), Use Silverware (A9), and Write on Board (A10).

To maintain consistency with existing approaches, we employed a five-fold leave-one-person-out (LOPO) evaluation strategy. In this method, one individual's samples are used exclusively for testing, while the remaining subjects' samples form the training set. Results are averaged across all five folds. For MHI construction, 30–40 frames were considered, and key-point trajectories were tracked over 30 frames. Table 1 summarizes the comparative performance of our approach, which achieved an overall accuracy of 95.4% on the URADL dataset.

We further evaluated different codebook sizes and selected a size of 1000 as an optimal balance between computational efficiency and recognition performance. The confusion matrix in Fig. 4 highlights misclassifications involving similar actions, such as "eat banana," "eat snack," "dial phone," and "answer phone."

Table 1. Comparison on existing methods (URADL Dataset)

Method	Classification Accuracy
Messing R *et* al. [9]	89.3%
Wang H. *et* al. [47]	92.7%
Yan Y. *et* al. [49]	88.1%
Avgerinakis K. *et* al. [50]	94.4%
Selmi *et* al. [51]	93.3%
Proposed method	95.4%

Fig. 4. Confusion matrix of codebook size 1000 codeword's

In contrast, actions including "chop banana", "look up in phone book", and "write on whiteboard" were recognized with 100% accuracy.

4.2. RESULTS ON KTH DATASET

The 600 video clips in the publicly available KTH dataset represent six different actions: walking (Walk), jogging (Jog), hand waving (HW), boxing (Box), and hand clapping (HC). A total of twenty-five people performed these actions in both indoor and outdoor environments. We used a conventional evaluation configuration to assess our proposed method's performance [7, 16, 42, 47]. In this setting, the first 16 subjects were used for training, while the remaining 9 subjects were reserved for testing. For MHI computation, 15–25 frames were utilized, and key-point trajectories were tracked over 15 consecutive frames.

Table 2 presents a comparative analysis with existing methods, showing that our proposed approach, which employs the composite feature descriptor, performs among the best on the KTH dataset. Our method achieved an accuracy of 95.83% using a codebook size of 1000 visual words. Furthermore, it was observed that increasing the codebook size beyond this value did not yield any significant improvement in accuracy.

Fig. 5 displays the action recognition confusion matrix for the KTH dataset, where similar actions such as boxing, hand clapping, jogging, and running sometimes exhibit misclassification. Our proposed technique achieved an average recognition rate of 95.83%, demonstrating its robustness on this benchmark dataset.

Table 2. Comparison of existing methods on KTH Dataset

Method	Classification accuracy
Laptev I. [16]	91.8%
Zhang Z. *et* al. [46]	93.5%
Zare A. *et* al. [52]	93.63%
Selmi M. *et* al. [51]	95.8%
Al-Berry *et* al. [7]	96%
Uddin M. A [53]	96.5%
Khan, M.A *et* al. [54]	97%
Garg A. *et* al [33]	96.24%
Khater S. *et* al.[55]	98.5%
Proposed Method	95.83%

Fig. 5. Confusion matrix for codebook size 1000 codeword's

4.3. RESULTS ON WEIZMANN DATASET

The The Weizmann dataset is a widely used benchmark for evaluating human action recognition methods [16]. It contains nine distinct actions performed by ten individuals. The actions include: Walk (A1), Run (A2), Jump Jack (A3), Bend (A4), Jump (A5), Jumping in Place (A6), Sideway Jump (A7), One-Hand Wave (A8), and Two-Handed Wave (A9). Table 3 presents a performance comparison using the Weizmann dataset alongside existing approaches. The results show that our proposed method achieves 100% classification accuracy using the combined feature descriptor, placing it among the top-performing methods on this dataset.

Table 3. Comparison of existing methods on Weizmann Dataset

Method	Classification accuracy
Gorelick *et* al. [11]	97.5%
Melfi *et* al. [56]	99.02%
Al-Berry *et* al. [7]	91.4%
Vishwakarma [57]	100%
Zare *et* al. [52]	100%
Garg *et* al. [33]	93.39%
Khater *et* al. [55]	99.2%
Proposed method	100%

Fig. 6 shows the confusion matrix for a codebook with 1000 codewords.

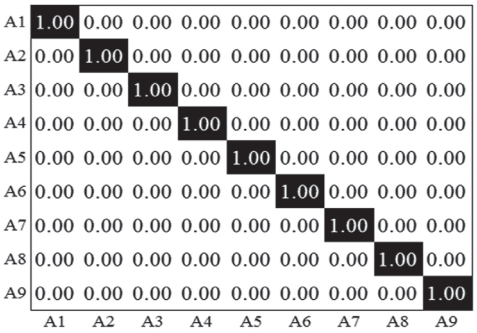


Fig. 6. confusion matrix for codebook size 1000 codeword's on Weizmann dataset

Although our framework, along with those proposed by Vishwakarma *et* al. [57] and Zare *et* al. [52], all reach perfect accuracy, their underlying approaches differ significantly. Vishwakarma *et* al. focus on modeling action dynamics using motion energy and temporal gradients. While this is highly effective for simple, single-person, and background-controlled datasets like Weizmann, their reliance on handcrafted motion information reduces robustness under camera motion, illumination variation, or multi-subject scenarios. Zare *et* al. employ a deep CNN-based spatiotemporal mapping model that learns compact features directly from raw frames. Despite its strength, such deep methods typically require large training datasets, greater computational resources, and may overfit when sample sizes are limited.

In contrast, our approach integrates global (MHI-HODD), local (intensity-HODD), and trajectory-shape descriptors into a unified Bag-of-Visual-Words (BoVW) and SVM framework. While achieving perfect accuracy on Weizmann, it also demonstrates superior robustness and generalization across more challenging datasets such as URADL (95.4%) and UCF101 (89%), making it both computationally efficient and interpretable.

4.4. RESULTS ON UCF 101 DATASET

To demonstrate the robustness and generalization capability of our proposed method, we evaluated it on the more challenging and complex UCF101 dataset, which consists of 101 action classes. Our approach performed competitively and achieved an accuracy of 89%.

4.5. SENSITIVITY ANALYSIS

Following parameters are considered for sensitivity analysis:

- Window size of HODD descriptor: **16×16, 32×32 and 64×64**
- Number of frames used to construct MHI-**15, 30 and 45 Frames**
- Trajectory tracking Length (frames): **15, 20, 25, 30 frames**
- Codebook Size : **500, 1000, 1500 and 2000**.

Parameter sensitivity analysis indicates that the performance of our HAR approach remains stable across a broad range of parameter settings. However, window size, number of orientation bins, and trajectory length emerge as the most influential factors affecting both accuracy and computational complexity. In particular, window sizes smaller than 32×32 or excessively large lead to a noticeable decline in recognition accuracy. The choice of 9 orientation bins offers the best balance between feature discriminative power and computational efficiency. Similarly, a trajectory length of approximately 20–25 frames yields optimal accuracy and robustness. The analysis confirms that our method is resilient to moderate deviations from these optimal values, demonstrating strong practical applicability across diverse real-world scenarios

5. CONCLUSION

This paper introduces an integrated framework that combines appearance and structural (spatial) features with motion features (temporal and shape of trajectories). The appearance and structural information of key points is effectively captured by our HODD descriptor. Through combining our proposed descriptor with MHI

and trajectory shape, we obtained more informative and discriminative spatial and temporal descriptor for key points. The trajectories shapes along the MHI exhibit greater robustness because of the predominant motion information along the moving object. The experimental outcomes unequivocally illustrate that the proposed method adeptly distinguishes between similar actions in the KTH dataset, such as walking, jogging, and running. The KTH, ADL (URADL), Weizmann and UCF 101 datasets yield classification rates of 95.83%, 95.4%, 100% and 89% respectively. Our forthcoming research endeavor's will concentrate on recognizing additional Activities of Daily Living (ADL).

REFERENCES

[1] R. Poppe, "A survey on vision based human action recognition, Image and vision computing", Image and Vision Computing, Vol. 28, 2010, pp. 976-990.

[2] M. Rodriguez, "CRAM: Compact representation of actions in movies", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13-18th June 2010.

[3] J. K. Aggarwal, M. Ryoo, "Human Activity Analysis A Review", ACM Computing Surveys, Vol. 43, 2011, pp. 1-43.

[4] Y. Wang, Y. Shi, G. Wei, "A novel local feature descriptor based on energy information for human activity recognition", Neurocomputing, Vol. 228, 2017, pp. 19-28.

[5] M. Uddin, J. B. Joolee, A. Alam, "Human Action Recognition Using Adaptive Local Motion Descriptor in Spark", IEEE Access, Vol. 5, 2017, pp. 21157-21167.

[6] X. Zhen, F. Zheng, L. Shao, X. Cao, "Supervised Local Descriptor Learning for Human Action Recognition", IEEE Transactions on Multimedia, Vol. 19, No. 9, 2017, pp. 2056-2065.

[7] M. N. Al-Berry, A. M. Salem, H. M. Ebeid, "Fusing directional wavelet local binary pattern and moments for human action recognition", IET Computer Vision, Vol.10, No. 2, 2016, pp. 153-162.

[8] S. Liu, J. Liu, T. Zhang, "Human action recognition in videos using hybrid features", Advances in Multimedia Modelling, Springer, 2010, pp. 411-421.

[9] R. Messing, C. Pal, H. Kautz, "Activity recognition using velocity histories of tracked key points", Proceedings of the IEEE International conference on Computer Vision, Kyoto, Japan, 29 September - 2 October 2009, pp. 104-111.

[10] C. Schuldt, I. Laptev, B. Caputo, "Recognizing human actions: a local SVM approach", Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 26 August 2004.

[11] L. Gorelick, M. Blank, E. Shechtman, M. Irani, "Actions as space-time shapes", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, No. 12, 2007, pp. 2247-2253.

[12] A. Bobick, J. W. Davis, "The recognition of human movement using temporal template", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, 2001, pp. 257-267.

[13] J. W. Davis, A. Bobick, "The representation and recognition of human movement using temporal templates", Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, PR, USA, 17-19 June 1997, pp. 928-934.

[14] D. M. Tsai, W. Chiu, M.H. Lee, "Optical motion history image (OF-MHI) for action recognition", Signal Image and Video Processing, Vol. 9, 2015, pp. 1897-1906.

[15] S. Islam, T. Qasim, M. Yasir, "Single- and two-person action recognition based on silhouette shape and optical point descriptors", Signal Image and Video Processing, Vol.12, No. 5, 2018, pp. 853-860.

[16] I. Laptev, "On space time interest points", International Journal of Computer Vision, Vol. 64, No. 3, 2005, pp. 107-123.

[17] H. Wang, C. Schmid, "Action recognition with improved trajectories", Proceedings of the International Conference on Computer Vision, Sydney, NSW, Australia, 1-8 December 2013, pp. 3551-3558.

[18] H. Wang, A. Kläser, C. Schmid, "Dense trajectories and motion boundary descriptors for action recognition", International Journal of Computer Vision, Vol. 103, No. 1, 2013, pp. 60-67.

[19] S. Singh, C. Arora, C. V. Jawahar, "Trajectory aligned features for first person action recognition", Pattern Recognition, Vol. 62, 2017, pp. 45-55.

[20] X. Wang, C. Qi, "Saliency-based dense trajectories for action recognition using low-rank matrix decomposition", Journal of Visual Communication and Image Representation, Vol. 41, 2016, pp. 361-374.

[21] R. Messing, C. Pal, H. Kautz, "Activity recognition using the velocity histories of tracked keypoints", Proceedings of IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September - 2 October 2009, pp. 104-111.

[22] C. Quan-Qi, Y. Zhang, "Cluster trees of improved trajectories for action recognition", Neurocomputing, Vol. 173, Part 2, 2016, pp. 364-372.

[23] P. Wang, W. Li, C. Li, "Action recognition based on joint trajectory maps with convolutional neural networks", Knowledge-Based Systems, Vol. 158, 2018, pp. 43-53.

[24] H. Arif, T. Ul-Hassan, F. Hussain, "Video representation by dense trajectories motion map applied to human activity recognition", International Journal of Computers and Applications, Vol. 42, 2018, pp. 474-484.

[25] A. Eison, C. V. Jiji, "Automated Video Analysis for Action recognition using descriptors derived from optical acceleration", Signal Image and Video Processing, Vol. 13, No. 5, 2019, pp. 915-922.

[26] A. M. Hamid, A. Zare, "Spatiotemporal wavelet correlogram for Human action recognition", International Journal of Multimedia Information Retrieval, Vol. 8, No. 3, 2019, pp. 167-180.

[27] M. Ahmad, S. W. Lee, "Recognizing human actions based on silhouette energy and global motion description", Proceedings of the 8th IEEE International Conference on Automatic Face & Gesture Recognition, Amsterdam, Netherlands, 17-19 September 2008, pp. 523-588.

[28] Y. Tian, L. Cao, Z. Liu, "Hierarchical filtered motion for action recognition in crowded videos", IEEE Transactions on Systems, Man, and Cybernetics: Part C, Vol. 45, 2012, pp. 313-323.

[29] D. Zhao, L. Shao, X. Zhen, "Combining appearance and structural features for human action recognition", Neurocomputing, Vol. 113, 2013, pp. 88-96.

[30] J. Yu, M. Jeon, W. Pedrycz, "Weighted feature trajectories and concatenated bag-of-features for action recognition", Neurocomputing, Vol. 131, 2014, pp. 200-207.

[31] C. L. Diogo, T. Hedi, P. David, "Learning features combination for human action recognition from skeleton sequences", Pattern Recognition Letters, Vol. 99, 2017, pp. 13-20.

[32] Z. Tong, Y. Song, J. Wang, L. Wang, "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training", Proceedings of the 36th Conference on Neural Information Processing Systems, New Orleans, LA USA, 8 November - 9 December 2022.

[33] A. Garg, S. Nigam, R. Singh, "Vision based Human Activity Recognition using Hybrid Deep Learning", Proceedings of the International Conference on Connected Systems & Intelligence, Trivandrum, India, 31 August - 2 September 2022, pp. 1-6.

[34] M. G. Morshed, A. S. Bhuiyan, H. Rahman, "Human Action Recognition: A Taxonomy-Based Survey", Sensors, Vol. 23, No. 4, 2023, pp. 2182-2204.

[35] A. Sánchez-Caballero, F. J. Muñoz, L. M. Bergasa, "Real-Time Human Action Recognition Using Raw Depth Video", Computer Vision and Image Understanding, Vol. 234, 2023, pp. 103729.

[36] Z. Hu, J. Wang, Y. Zhang, "Human-Centric Multimodal Fusion Network for Robust Action Recognition", IEEE Transactions on Multimedia, Vol. 26, 2024, pp. 4230-4244.

[37] A. Zhang, Q. Li, J. Xu, "Advancing Human Action Recognition with Vision-Language and Foundation Models", Pattern Recognition Letters, Vol. 179, 2025, pp. 20-33.

[38] M. Kamble, R. S. Bichkar, "A Hierarchical Framework for Video-Based Human Activity Recognition Using Body Part Interactions", International Journal of Electronics and Computer Engineering Systems, Vol. 14, No. 8, 2023, pp. 881-891.

[39] S. Abraham, R. K. James, "Significance of Handcrafted Features in Human Activity Recognition with Attention-Based RNN Models", International Journal of Electronics and Computer Engineering Systems, Vol. 14, No. 10, 2023, pp. 1151-1163.

[40] D. S. Korti, Z. Slimane, "Advanced Human Activ-

ity Recognition through Data Augmentation and Feature Concatenation of Micro-Doppler Signatures", International Journal of Electronics and Computer Engineering Systems, Vol. 14, No. 8, 2023, pp. 893-902.

[41] S. B. Bhorge, R. R. Manthalkar, "Recognition of Vision-Based activities of daily living using linear predictive coding of histogram of directional derivative", International Journal of Ambient Intelligence and Humanized Computing, Vol. 10, No. 1, 2017, pp. 199-214.

[42] B. D. Lucas, T. Kanade, "An iterative image registration technique with an application to stereo vision", Proceedings of the 7th international joint conference on Artificial intelligence, Vancouver, BC, Canada, 24 August 1981.

[43] M. A. Ahad, T. J. Kim, S. Ishikawa, "Motion History image: its variants and applications", Machine Vision and Applications, Vol. 23, No. 2, 2012, pp. 255-281.

[44] C. Harris, M. Stephens, "A combined corner and edge detector", Proceedings of the Alvey Vision Conference, 1988, pp. 189-192.

[45] N. Dalal, B. Triggs, "Histogram of oriented gradients for human detection", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20-25 June 2005.

[46] Z. Zhang, D. Tao, "Slow feature analysis for human action recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 46, 2012, pp. 1810-1818.

[47] H. Wang, A. Klaser, C. Schmid, "Action recognition by dense trajectories", Proceedings of the IEEE Inter. Conf. on Computer Vision & Pattern Recognition, Colorado Springs, CO, USA, 20-25 June 2011, pp. 3169-3176.

[48] S. Wu, V. Pham, "Speeding up multi-SVMs through modified working set selection", International Journal of Computers and Applications. Vol. 44, No. 5, 2020, pp. 426-432.

[49] Y. Yan, R. Elisa, R. Negar, "It's all about habits: Exploiting multi-task clustering for activities of daily living analysis", Proceedings of the IEEE International Conference on Image Processing, Paris, France, 27-30 October 2014 pp. 1071-1075.

[50] K. Avgerinakis, A. Briassouli, K. Loannis, "Activities of daily living recognition using optimal trajectories from motion boundaries", Journal of Ambient Intelligence and Smart Environments, Vol. 7, No. 6, 2015, pp. 817-834.

[51] M. Selmi, M. A. El-Yacoubi, B. Dorrizi, "Two layer discriminative model for human activity recognition", IET Computer Vision, Vol.10, No. 4, 2016, pp. 273-278.

[52] A. Zare, H. A. Moghaddam A. Sharifi, "A. Video spatiotemporal mapping for human action recognition by convolutional neural network", Pattern Analysis Applications, Vol. 23, 2020, pp. 265-279.

[53] M. A. Uddin, Y. Lee, "Feature Fusion of Deep Spatial Features and Handcrafted Spatiotemporal Features for Human Action Recognition", Sensors, Vol. 19, 2019.

[54] M. A. Khan, K. Javed, S. A. Khan, "Human action recognition using fusion of multiview and deep features: an application to video surveillance", Multimedia Tools and Applications, Vol. 83, 2020, pp. 14885-14911.

[55] S. Khater, M. Hadhoud, M. B. Fayek., "A novel human activity recognition architecture: using residual inception Conv LSTM layer", Journal of Engineering and Applied Science, Vol. 69, No. 45, 2022.

[56] R. Melfi, S. Kondra, A. Petrosino, "Human activity modeling by spatio-temporal textural appearance", Pattern Recognition, Vol. 34, 2013, pp. 1990-1994.

[57] D. K. Vishwakarma, K. Rajiv, D. Ahish, "A Proposed unified framework for the recognition of human activity by exploring the characteristics of action dynamics", Robotics and Autonomous Systems, Vol. 77, 2016, pp. 25-38.