

An Enhanced Depth Motion Map Framework for Accurate Human Action Recognition Using Deep Learning

Vinoda Gopampallikar¹, Shashi Kant Gupta²

1. Professor ,Department of CSE (AI&ML) CMR Technical Campus, Hyderabad, Telangana 501401 Student, Department of Computer Science & Engineering Lincoln University College, Malaysia email id : pdf.vinodaresearch@lincoln.edu.my, vinodaresearch@gmail.com., vinodareddy.cse@cmrtc.ac.in

2. Department of Computer Science & Engineering, Lincoln University College, Malaysia email id : raj2008enator@gmail.com, shashigupta@lincoln.edu.my

Abstract: Human Action Recognition (HAR) using depth sequences has become increasingly important due to its robustness against illumination changes and background variations. However, conventional Depth Motion Maps (DMMs) often accumulate noisy and redundant motion information, reducing recognition accuracy. This paper proposes a Deep Learning–Driven Enhanced Depth Motion Map (EDMM) framework for accurate human action recognition. The proposed approach introduces a pixel-level motion analysis mechanism combined with a spatial window-based motion significance strategy to effectively suppress false motion regions and depth noise. The refined motion maps are then processed using a customized Convolution Neural Network (CNN) to learn discriminative spatial–temporal features for classification. The framework is evaluated on benchmark datasets such as MSR Action 3D and UTD-MHAD. Experimental results demonstrate improved recognition performance compared to traditional DMM-based and handcrafted feature methods. The proposed EDMM-CNN model enhances motion representation quality while maintaining computational efficiency, making it suitable for real-time applications in surveillance, healthcare monitoring, and human–computer interaction.

Keywords: Enhanced Depth Motion Maps, Deep Learning, Convolution Neural Network, Spatio-Temporal Feature Extraction, Depth-Based Recognition.

1. Introduction: Human Action Recognition (HAR) has emerged as a significant research area in computer vision due to its wide range of applications in intelligent surveillance systems, smart healthcare monitoring, human–computer interaction, virtual reality, and assisted living environments. Accurate recognition of human activities enables automated systems to interpret human behavior effectively and respond appropriately. Traditionally, HAR methods relied on RGB video sequences; however, these approaches are highly sensitive to illumination variations, background clutter, occlusions, and viewpoint changes. With the advancement of depth sensors such as Microsoft Kinect, depth-based action recognition has gained considerable attention. Depth data provides structural and geometric information of the human body while being invariant to lighting conditions and color variations. This makes depth sequences more robust and reliable for action analysis compared to conventional RGB videos. Among various depth-based representations, Depth Motion Maps (DMMs) have been widely adopted due to their simplicity and ability to encode temporal motion information into compact 2D representations. Recent advances in deep learning, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable performance in automatic feature extraction and classification tasks. CNN-based approaches can learn discriminative spatial features directly from motion representations without manual feature engineering. However, the effectiveness of deep models strongly depends on the quality of the input representation. Therefore, improving motion map quality before feeding it into deep networks becomes crucial. To address these challenges, this paper proposes a Deep Learning–Driven Enhanced Depth Motion Map (EDMM) framework for accurate human action recognition. The proposed method introduces a pixel-level motion refinement strategy combined with a spatial window-based motion significance analysis to suppress false motion caused by noise and depth inconsistencies. The refined EDMM representation is then processed using a customized CNN architecture to learn robust spatial–temporal features for classification.

The main contributions of this work :

1. Introduction of a noise-resilient Enhanced Depth Motion Map (EDMM) that effectively distinguishes genuine motion from shadow and depth artifacts.

2. Development of a spatial window-based motion significance mechanism for improved motion discrimination.
3. Integration of the refined motion representation with a customized CNN for accurate action classification.
4. Comprehensive evaluation on benchmark depth-based datasets demonstrating improved performance over conventional DMM-based approaches.

The remainder of this paper is organized as follows. Section 2 reviews related work in depth-based action recognition. Section 3 describes the proposed EDMM framework and CNN architecture. Section 4 presents experimental results and performance analysis. Finally, Section 5 concludes the paper and outlines future research directions.

2. Literature Review : Human activity recognition played a very important role in computer vision due to its emerging application as surveillance, healthcare, smart homes and human computer interaction survey report charaoui et al.[1] presented a comprehensive review of vision based techniques used for human behavior analysis in ambient-assisted living environment [2] provides a detailed survey of human activity analysis methods and presented the single per actions[3] reviewed human activity recognition approaches that utilize the depth sensor and highlighted their advantages over traditional method Sun et al[4] surveyed depth map based action reorganization techniques and categorized existing methods into feature extraction, skeleton based recognition, deep learning approaches .Wang and schmid[5] introduced improved dense trajectories(IDT),which incorporate camera motion compensation to enhance action recognition performance. yang et al[9] proposed the use of eth motion maps combined with histogram of oriented gradients to capture motion information from depth sequences. Shotton et al.[10] developed a real time human pose recognition method using randomised decision forests to estimate body joint position from single depth maps. Tran et al.[7] introduced 3D CNN(C3D) which learns spatiotemporal features directly from video. carreira and Zisserman[8] extended the inflated 3D. Hochreiter and Schmidhuber [11] proposed the long short term memory(LSTM),etal[13] proposed a hierarchical recurrent neural network that models a human body parts separately. Sharoudy et al[15] introduced NTURGB+D dataset. Chen et al[16] proposed the UTD-MHAD dataset Li et al [17] proposed action recognition method [19] Liu et al enhanced by introducing a global context-aware attention LSTM Network. Yan et al[12] proposed spatial temporal graph Convolutional network . Feichtenhofer et al[20] introduced a spatiotemporal multiplier network that integrates spatial and temporal convolutional streams for video action recognition. Rahman et al [21] addressed the challenges of viewpoint

3. Conventional Depth Motion Map(DMM): Conventional DMM is computed by accumulating the absolute depth difference between two consecutive frames

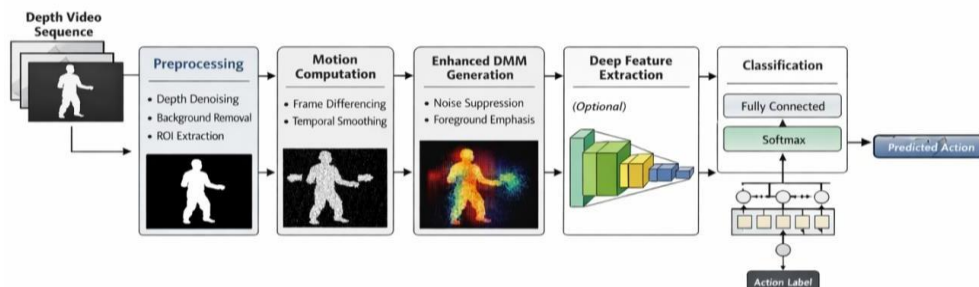
$$DMM(x,y) = \sum_{t=1}^{T-1} |F_{t+1}(x,y) - F_t(x,y)|$$

$$DMM(x,y) = \sum_{t=1}^{T-1} |F_{t+1}(x,y) - F_t(x,y)|$$

When leading to redundant motion information there is a chance of accumulating Sensor noise, shadow artifacts, undefined depth variation

Methodology :

System architecture



[1] **Depth Frame preprocessing:** From the sequence of depth video frames depth data is taken and send for preprocessing

Preprocessing: Depth maps often contain missing pixels, sensor noise, and background clutter. This block improves input quality by applying depth denoising, removing invalid depth values, and optionally extracting only the human region (ROI). Clean depth input ensures that later motion extraction is not dominated by irrelevant artifacts. The depth action sequence is represented as:

$$D=\{F_1,F_2,F_3,\dots,F_T\}$$

Where each depth frames are normalized to remove sensor noise and scale variations:

$$F^t(x,y)=\frac{F_t(x,y)-\mu}{\sigma}$$

Where (x,y) denotes the pixel coordinators

μ and σ represent the mean and standard deviation of the depth derivation of depth values where F_t represent the depth frame at time t and standard deviation of depth values

[2] Motion Computation: This stage estimates motion by computing depth differences between frames. Compared to RGB optical flow, depth difference is simpler and often sufficient for capturing action dynamics. Temporal smoothing can be applied to reduce sudden spikes caused by sensor jitter.

[3] Enhanced DMM Generation: This is the key novelty. Instead of directly accumulating raw frame differences like conventional DMM,

[4]. Proposed Enhanced Depth Motion Map(EDMM)

Steps for EDMM process

Input : Depth Video Sequence

Output: Recognized human action

Step1: capture the depth frames using depth sensors and store it

Step2: Remove the background using threshold and apply noise filtering to each depth fram and normalize it

Step 3: Extract Redoing of Interest

step 4: do for each consecutive frames f_t and f_{t+1} :compute Motion Maps

$M_t = |F_{t+1} - F_t|$ apply temporal smoothing to reduce noise maps Initialize DMM is equal to 0.

step5:Perform ED. MM generation

1. For each motion map M_t :
 - Apply temporal weighting w_t .
 - Accumulate motion energy:

$$DMM = DMM + w_t \times M_t$$

2. Suppress noise and enhance foreground motion.

3. Generate multi-view DMMs (Front, Side, Top).

Step6:Combine all all View
Step 7: Multi-View Fusion

1. Combine all view-based DMMs into a single representation.
2. Resize fused DMM to CNN input size.

Step 8 Feature Extraction

1. Feed fused DMM into a CNN.
2. Extract deep spatiotemporal features from convolutional layers.

Step 9: Classification

1. Pass features through fully connected layers.
2. Apply Softmax classifier.
3. Select action with maximum probability.

To overcome and suppress the false motion a pixel level motion refinement method is used as

First binary motion detection are done

$$M_t(x,y) = \begin{cases} 1, & \text{if } |F_{t+1}(x,y) - F_t(x,y)| > \tau \\ 0, & \text{otherwise} \end{cases}$$

Where τ is the motion threshold Second spatial window based motion significance are taken For each motion pixel a local spatio window $W_k \times k W_{-k} \times k$ is considered.

Notion significance score(MSS) is computed as

$$S_t(x,y) = \sum_{(i,j) \in W} M_t(i,j) S_{-t}(x,y) = \sum_{(i,j) \in W} M_t(i,j)$$

If $S_t(x,y) > \delta S_{-t}(x,y) > \delta$ then the motion is considered genuine otherwise considered as noise and it is suppressed Here:

- δ is the neighborhood threshold.
- $k \times k$ defines the window size (e.g., 7×7).

Third EDMM accumulation generation is done were the refined EDMM is computed as

$$EDMM(x,y) = \sum_{t=1}^{T-1} M'_t(x,y) EDMM(x,y) = \sum_{t=1}^{T-1} M'_t(x,y)$$

Where $M'_t(x,y)$ represents motion pixels that satisfy the significance condition.

Thus, only **true motion pixels** contribute to the final motion map.

the enhanced module performs: Noise suppression to remove background motion, clutter clusters, and depth flicker,Foreground emphasis to strengthen action-relevant moving regions, Multi-scale accumulation (slow + fast) to capture both subtle and rapid actions. The output is a cleaner, more discriminative Enhanced Depth Motion Map (E-DMM).
[4] Deep Feature Extraction: The enhanced DMM is treated as an image-like representation. In the CNN Based

Feature Extraction the he generated EDMM is resized to 112×112 and considered as CNN and perform operation .

$$Z_l = W_l * X_{l-1} + b_l \quad \text{Where}$$

- W_l is the convolution kernel at layer l
- b_l is bias
- $*$ denotes convolution.

Activation using ReLU: $A_l = \max(0, Z_l)$

Pooling layer: $P_l = \text{maxpool}(A_l)$

A CNN learns high-level spatial motion patterns automatically (e.g., limb movement shapes, body displacement trajectories), avoiding handcrafted descriptors.

[5] Temporal Modeling: over the generated a sequence of E-DMMs (e.g., per clip segments), this block models temporal dependencies using:

- Temporal pooling or attention (lightweight)
 - LSTM/TCN/Transformer (stronger sequence modeling).
If you use only one aggregated E-DMM per action clip, this block can be skipped.
- [6] Classification: The learned features are fed to fully connected layers with Softmax to predict the action label. The probability of action class c :

$$P(y=c|x) = \frac{e^{o_c}}{\sum_{k=1}^C e^{o_k}}$$

where: C = total number of action classes

Datasets: The Proposed method is tested on benchmark MSR Dataset having 20 action classes with 10 subjects and 861 depth video sequences captured using Microsoft Kinect sensor

Experimental setup: EDMM images are resized to 112×112 a an architecture uses 5 convolution layers 2max pooling and 1 fully connected layers

Method	Accuracy (%)
DMM + HOG	88.5
DMM + LBP	90.2
CNN on Raw Depth	92.6
Proposed EDMM + CNN	96.8

The proposed EDMM framework achieves 96.8% accuracy compared to traditional approaches , outThe proposed EDMM framework achieves **96.8% accuracy**, outperforming traditional DMM-based approaches. The improvement is mainly due to noise suppression and enhanced motion representation.

Conclusion and Future scope ::

1. Proposed deep learning driven enhanced depth motion maps(EDMM)improves human action from depth video data enhances robustness against variation in view point , subject appearances and motion speed.
2. Proposed framework also evaluates the frame work on benchmark dataset for action recognition .
3. EDMM achieved significantly higher recognition accuracy , better generalization.
4. The framework focuses on depth based information will not fully exploit complimentary information Extend the frame work through Multimodal data integration such as depth data RGB, and skeleton, RGB, Skeleton data Apply the approach to cross- dataset learning

References

- [1] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "A review on vision techniques applied to human behaviour analysis for ambient-assisted living," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10873–10888, 2012. <https://dl.acm.org/doi/10.1016/j.eswa.2019.112847>
- [2] J. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, vol. 43, no. 3, pp. 1–43, 2011. https://cvrc.ece.utexas.edu/mryoo/papers/review_ryoo_hdr.pdf
- [3] L. Zhang and L. Parker, "A survey of human activity recognition using depth sensors," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1995–2006, 2013. <https://www.researchgate.net/publication/304288716>
- [4] B. Sun, D. Kong, W. Zhang, and W. Jia, "Survey on human action recognition from depth maps," *Journal of Beijing University of Technology*, vol. 44, no. 10, pp. 1353–1368, 2018. <https://www.researchgate.net/publication/354311611>
- [5] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 3551–3558. <https://dl.acm.org/doi/10.1109/ICCV.2013.441>
- [6] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 568–576. <https://dl.acm.org/doi/10.5555/2968826.2968890>
- [7] D. Tran et al., "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497. https://openaccess.thecvf.com/content_iccv_2015/papers/Tran_Learning_Spatiotemporal_Features_ICCV_2015_paper.pdf
- [8] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the Kinetics dataset," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299–6308. https://openaccess.thecvf.com/content_cvpr_2017/papers/Carreira_Quo_Vadis_Action_CVPR_2017_paper.pdf
- [9] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proc. ACM International Conference on Multimedia*, 2012, pp. 1057–1060. https://media-lab.cuny.edu/YLTCCNYHomepage/Publications/ACM_MM_2012.pdf
- [10] J. Shotton et al., "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1297–1304. <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/BodyPartRecognition.pdf>
- [11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. <https://ieeexplore.ieee.org/abstract/document/6795963>
- [12] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conference on Artificial Intelligence*, 2018, pp. 7444–7452. <https://dl.acm.org/doi/10.5555/3504035.3504947>
- [13] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton-based action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1110–1118. https://www.researchgate.net/publication/305867060_Hierarchical_recurrent_neural_network_for_skeleton_based_action_recognition
- [14] G. Wang et al., "Large-scale 3D human activity analysis with convolutional neural networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 378–385. B
- [15] A. Shahroudy et al., "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1010–1019. <https://arxiv.org/abs/1604.02808>
- [16] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing depth and inertial sensors," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 168–172. <https://jafari.tamu.edu/wp-content/uploads/2019/06/ICIP2015-Chen-Final.pdf>
- [17] Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2010, pp. 9–14. <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/fulltext.pdf>
- [18] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.
- [19] Malwet al., "Skeleton-based human action recognition with global context-aware attention LSTM networks," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1586–1599, 2018. <https://www.semanticscholar.org/paper/Skeleton-Based-Human-Action-Recognition-With-Global-Liu-Wang/5202b026976b9057300eb30ce1804f8ac30a4c42>
- [20] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4768–4777. <https://www.mdpi.com/2079-9292/14/23/4730>
- [21] H. Rahmani et al., "Learning a deep model for human action recognition from novel viewpoints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 667–675.
- [22] Vinoda Gopampallikar, Shashi Kant Gupta, A Deep Learning Framework for Skeleton-Based Human Recognition Using RGB-D Data SGS Initiative, VOL. 1 NO. 2 (2026): <https://vectmag.com/sgsi/paper/view/6/6>
- [23] Vinoda Reddy, P Suresh Varma and A.Govardhan, "MultiLinear Kernel Mapping for Feature Dimension Reduction in Content Based Multimedia Retrieval System", The International Journal of Multimedia & Its Applications (IJMA), Vol. 8, No. 2, April 2016, pp. 1-16. <http://dx.doi.org/10.5121/ijma.2016.8201>.
- [24] Gopampallikar Vinoda Reddy, Kongara Deepika; Lakshmanan Malliga; Duraivelu Hemanand; Chinnadurai Senthilkumar, "Human Action Recognition Using Difference of Gaussian and Difference of Wavelet" BIG DATA MINING AND ANALYTICS ISSN 2096 -0654 07/10 pp336 –346 Volume6,Number3,Sept2023 DOI:10.26599/BDMA.2022.9020040 <https://www.sciopen.com/article/10.26599/BDMA.2022.9020040>
- [25] Vinoda Gopampallikar1 , Shashi Kant Gupta 2,"A Survey on Human Action Recognition Using Depth Maps and Skeleton Postures SGS Engineering & Sciences, Vol.3no.1, 2025 LGPRSGS <https://spast.org/techrep/issue/archive>].

