Indian Journal of Computer Science and Technology
Volume 3, Issue2 (May-August 2024), PP: 207-212.
www.indjcst.com
ISSN No: 2583-5300

State-of-the-Art in Human Locomotion Action Recognition: A Review

Paras Jain[1], Meenakshi Arora[2], Rohini Sharma[3]
[1]*P.G. Student, Department of CSE, Sat Kabir Institute of Technology and Management, Haryana, India.*
[2]*Assistant Professor, Department of CSE, Sat Kabir Institute of Technology and Management, Haryana, India.*
[3]*Assistant Professor, Government P.G. College for Women, Rohtak, Haryana, India.*

To Cite this Article: *Paras Jain[1], Meenakshi Arora[2] Rohini Sharma[3], "State-Of-The-Art in Human Locomotion Action Recognition: A Review", Indian Journal of Computer Science and Technology, Volume 03, Issue 02 (May-August 2024), PP: 207-212.*

Abstract: *Human action recognition is a vital area of research in computer vision and machine learning, with applications spanning surveillance, healthcare, sports analysis, and human-computer interaction. This review presents a comprehensive overview of various human action recognition methods, highlighting their distinctive approaches and contributions to the field. We categorize these methods into segmentation-based, handcraft feature extraction, shape-based, motion-based, local binary pattern, and fuzzy logic approaches. Segmentation techniques focus on dividing the video into meaningful segments to isolate actions. Handcraft feature extraction involves manually designing features that capture relevant aspects of human motion. Shape-based methods analyze the silhouette or contour of the human body to identify actions, while motion-based methods focus on the dynamics of movement over time. Local binary pattern techniques leverage* texture information for action recognition. Lastly, fuzzy logic approaches incorporate uncertainty handling and approximate reasoning to improve recognition accuracy. This review aims to provide insights into the strengths and limitations of each method, guiding future research towards more robust and efficient human action recognition systems.

Keywords: *Human action recognition, feature extraction, Shape-based methods.*

I.INTRODUCTION

There are several possible uses for Human Action Recognition (HAR). Its goal is to identify a person's movements based on visual information or sensors. HAR techniques can be divided into three groups: multi-modal, non-visual sensor-based, and visual sensor-based[1]. The form of the felt data is the primary distinction between the visual and other categories. Some sources record the visual data as 1D signals, whereas others record the data as 2D, 3D, or video pictures [2]. Wearable technology has advanced over the past few years, with the development of smartwatches, fitness bands, and smartphones. These are outfitted with tiny, non-visual sensors as well as communication and processing power. Additionally, their very inexpensive cost has helped to provide new opportunities due to their widespread usage. These consist of disease prevention, rehabilitation training, and health surveillance[2].

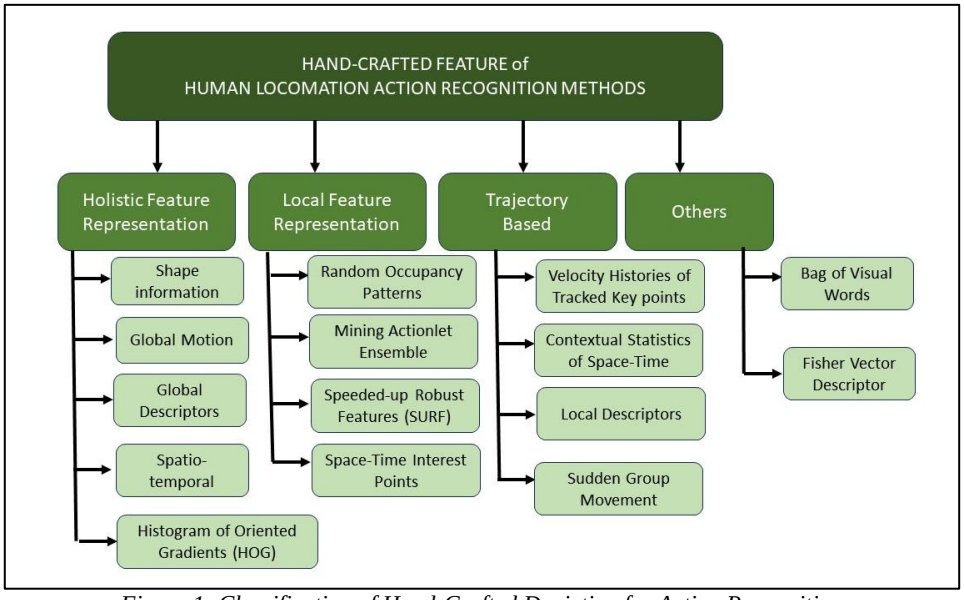

Figure 1: Classification of Hand-Crafted Depiction for Action Recognition

Simultaneously, among the most popular and hot topics in computer vision research are visual sensor-based approaches for human action identification. Applications include content-based video search, intelligent video surveillance, contextual assisted living, human-robot interaction, and human-computer interaction[3]. The recognition system is trained to discern between activities performed in a scene in each of those apps[4]. Based on that inference, it might also make some decisions or carry out additional processing. Based on the intricacy of human actions, action recognition systems can be divided into four groups. Primitive [5], solitary [6], interaction [7], and group [10] actions recognition are examples of this. Basic movements of human body parts, such as "lifting a hand" and "bending," are referred to as basic actions. Figure 1 shows basic classification of Hand-Crafted Attributes Depiction for Action Recognition.

II.RESEARCH BACKGROUND

The field of human action recognition along with associated approaches has seen the publication of numerous significant survey and review publications. Nevertheless, published reviews typically become outdated. This is why, although being a difficult and demanding undertaking, producing an updated study on human action recognition is greatly needed. This review presents the latest techniques for vision-based human action identification, along with debates, analysis, and comparisons. Popular benchmark datasets, important applications, and methodologies based on deep learning and handcrafted techniques are presented. Additionally, this article examined various recognition model designs, such as hybrid, modalities-based, and view-invariant based systems**.**

The majority of approaches presume that tasks are carried out from a set point of view. But in a genuine scenario, the person's position and posture differ significantly depending on the angle at which the action is photographed. Additionally, different views appear with varying motion patterns, which further complicates the detection of an action. [8] approaches this problem by using numerous camera information to train a classifier. In order to obtain a view-invariant representation, a 3D body posture model for action recognition was also created, as shown in [9]. By employing cylindrical coordinate systems and the Fourier transform, researchers attempt to make use of view-invariant features space [10]. The majority of multi-view datasets, according to researchers [11], contain a homogeneous or unchanging background.

Hand-Crafted Feature Representation for Action Recognition

Human action recognition using hand-crafted features has been a prominent approach in the early stages of this research field. This method involves manually designing and extracting features that are expected to be relevant for distinguishing different actions. Here, we discuss two major categories: holistic feature representation-based methods and hybrid methods based on shape and global motion information.

Holistic Feature Depiction Based Methods

Approaches that rely on holistic representation of features handle Regions of Interest (ROIs) holistically, utilizing every pixel to generate descriptors. Person identification and descriptor computing are the two phases that holistic based approaches often involve in order to recognize an action. In holistic approaches, the entire human body is seen as a representation of an activity, with no need to localize specific bodily parts. According to the data utilized for the recognition issue, holistic approaches can be divided into two groups:

Shape and Global Information Based Approaches

Below is a comparison of different shape information-based methods for human action recognition, categorized into silhouette information, color and location information, RGB-D information, Histogram of Oriented Gradients, and space-time-based methods.

Class	Technique	Approach	Strengths	Weaknesses	References
Silhouette Information	Vision-based human motion analysis	Summary of vision-based approaches by means of silhouettes	Thorough review, finds crucial approaches	Absence of implementation specifics, general summary	[12]
	Human shape-motion analysis in athletics videos	employs a transferable belief model to analyze shape and motion	Integrates form and action, making it effective for sports	Particular to sports, could not apply generally	[13]
	Temporal templates	Motion Energy Images (MEI) and Motion History Images (MHI) are used.	Time-based data collection and straightforward application	little spatial information and susceptible to noise and obstructions	[14]
Color and Location Information	Event detection in crowded videos	uses location and color characteristics in	Sturdy feature extraction that works well in	Expensive computing costs	[15]

		congested settings to detect occurrences	cluttered environments	and lighting sensitivity	
RGB-D Information	Color-Depth video database for daily activity recognition	Combines RGB and depth data for activity recognition	uses depth data and is resilient to changes in shape	depth sensors are needed, and data processing is expensive.	[16]
	Action recognition based on a bag of 3D points	makes use of 3D point cloud data to identify actions	combines robust description and 3D spatial information.	heavy computing and the need for depth sensors	[17]
	Action recognition using depth motion maps and local binary patterns (LBP)	integrate maps of depth motion with (LBP)	Reliable depth-based features that are resistant to occlusions	High processing cost and sensitivity to the quality of the depth map	[18]
Histogram of Oriented Gradients (HOG)	Human action recognition using depth motion map and KECA	employs KECA and a temporal cascading stack of depth motion map	Good motion depiction and a hierarchical framework	Exorbitant processing expenses and intricate feature extraction	[19]
Space-Time Based	Space-time occupancy patterns (STOP)	examines occupancy variations in space and time using depth map sequences	records 3D spatiotemporal data and is resistant to changing viewpoints.	high processing costs and the need for depth sensors	[20]
	HON4D: Histogram of oriented 4D normals	Recognizes activity from depth sequences using 4D normals	Thick spatiotemporal characteristics and strong identification	Expensive computing costs and depth quality sensitivity	[21]
Shape and Global Motion Information (Hybrid Approach)	DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification	use discriminative latent dirichlet assignment to learn categorization and reduce dimensionality.	Efficient dimensionality reduction increases the precision of categorization	intricate model requiring a lot of computing power	[22]
	Acknowledging Movement from a Distance	uses global motion characteristics to identify distant activities	robust to changes in perspective and scale, useful for remote actions	cognizant of background congestion and occlusions	[23]
	Effective Visual Event Recognition Employing Volumetric Characteristics	makes use of volumetric properties to effectively detect visual events	Strong detection and high feature extraction efficiency	High processing costs and huge memory needs when dealing with volumetric data	[24]

Local Feature Depictions Based Approaches

When it comes to RGB-based video, local feature-based techniques have a tendency to capture distinctive characteristics locally within a frame without requiring human identification or segmentation. Numerous recognition system applications, including action recognition, object identification, and scene recognition, have effectively used local feature-based approaches. Important shape and motion features for a particular region in a video can be captured by local capture-based techniques. Generally speaking, local feature-based techniques involve two stages: descriptor computing and point of interest (POI) detection. Interest spots in image processing are locations where there is a localized change in picture intensity.

Technique	Approach	Strengths	Weaknesses	References
Random Occupancy Patterns	uses random occupancy patterns to identify actions based on depth information	Properly captures spatial occupancy knowledge, appropriate for depth data	sensitive to noise in detailed data and expensive to compute	[25]
Mining Actionlet Ensemble for Action Recognition with Depth Cameras	mines depth camera actionlet ensemble for reliable action identification	Efficient for handling occlusions in depth camera data	big training datasets and a high processing load are necessary	[26]
Speeded-up Robust Features (SURF)	extracts strong local characteristics for action recognition using SURF.	Rapid feature extraction, scalability, and rotation	Misses small details in the motions and is less effective when there is a lot of motion blur	[14]
Space-Time Interest Points	finds interest locations in space and time to record local characteristics for action recognition.	Capable of recording motion dynamics and resilient to changes in speed	significant computational cost and noise and occlusion sensitivity	[15]

Table 2: Local Feature Depictions Based Approaches for Human Action Recognition

Trajectories Based Approaches

It has been asserted by numerous academics that the temporal and spatial domains in video have distinct properties. Therefore, it is not appropriate to detect points of interest in a 3D spatiotemporal domain. As a result, tracking identified locations of interest throughout the temporal domain has been a common practice in research. Next, the descriptors for video representation are frequently computed using the overall size of the trajectory points.

Technique	Approach	Strengths	Weaknesses	References
Activity Recognition Using the Velocity Histories of Tracked Keypoints	Utilizes velocity histories of tracked key points for recognizing activities	Adaptable to changes in direction and speed, accurately represents motion dynamics	sensitive to noise in detailed data and expensive to compute	[27]
Contextual Statistics of Space-Time Ordered Features for Human Action Recognition	utilizes space-time ordered characteristics' situational statistics to identify actions.	gathers contextual data well, robust to modifications in actions	High processing costs and a need for sizable training datasets	[28]
Local Descriptors for Spatio-Temporal Recognition	uses spatial-temporal data captured by local descriptors for motion analysis	Capable of capturing the dynamics of local motion, resilient to changes in speed	cognizant of occlusions and noise, theoretically costly	[29]
Towards Unsupervised Sudden Group Movement Discovery for Video Surveillance	concentrates on the unsupervised detection of abrupt group movements for monitoring	Identifying group movements effectively makes it appropriate for surveillance purposes.	restricted to collective acts; may overlook individual acts	[30]

Table 3: Trajectories Based Approaches for Human Action Recognition

Other Feature Depictions Approaches

Class	Technique	Approach	Strengths	Weaknesses	References
Bag of Words (BOW)	Visual Categorization with Bags of Keypoints	transforms local characteristics into a visual word histogram, allowing keypoint-based action identification.	Simple and effective, capable of handling a wide range of jobs, and simple to use.	depends on the quantity and caliber of the visual vocabulary; loses temporal and spatial information.	[31]
Fisher Kernels on Visual Vocabularies for Image Categorization	Fisher Kernels on Visual Vocabularies for Image Categorizatio	uses Gaussian Mixture Models to represent data and encodes higher-order statistics of the features to improve the BOW model.	gathers more precise data than BOW, performs better across a wide range of activities, and is resilient to change.	computationally demanding, more difficult to execute, and requires careful parameter adjustment.	[32]

Table 4: Bag of Words (BOW) and Fisher Vector Descriptors Approaches for Human Action Recognition

III.CONCLUSION

In this review, we compared several approaches for human action recognition (HAR), including holistic feature representation, local feature representation, trajectory-based methods, Bag of Words (BoW), and Fisher Vector descriptors. Each of these approaches offers unique strengths and weaknesses, making them suitable for different applications and scenarios.

References

1. *S. Ranasinghe, F. Al Machot, and H. C. Mayr, "A review on applications of activity recognition systems with regard to performance and evaluation," Int. J. Distrib. Sens. Networks, vol. 12, no. 8, p. 1550147716665520, 2016.*
2. *T. Sztyler, H. Stuckenschmidt, and W. Petrich, "Position-aware activity recognition with wearable devices," Pervasive Mob. Comput., vol. 38, pp. 281–295, 2017.*
3. *R. S. Monisa Nazir, Shalini Bhadola, Kirti Bhaia, "A Complete Analysis of Human Action Recognition Procedures," Int. J. Trend Sci. Res. Dev., vol. 6, no. 5, pp. 593–597, 2022.*
4. *R. S. Monisa Nazir, Kirti Bhatia, Shalini Bhadola, "Spatio-Temporal and Support Vector Machine Based Human Action Detection," Int. J. Multidiscip. Res. Sci. Eng. Technol. Manag., vol. 9, no. 7, pp. 1499–1505, 2022.*
5. *B. K. Chakraborty, D. Sarma, M. K. Bhuyan, and K. F. MacDorman, "Review of constraints on vision-based gesture recognition for human–computer interaction," IET Comput. Vis., vol. 12, no. 1, pp. 3–15, 2018.*
6. *D. Das Dawn and S. H. Shaikh, "A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector," Vis. Comput., vol. 32, pp. 289–306, 2016.*
7. *M. Meng, H. Drira, and J. Boonaert, "Distances evolution analysis for online and off-line human object interaction recognition," Image Vis. Comput., vol. 70, pp. 32–45, 2018.*
8. *B. Chakraborty, O. Rudovic, and J. Gonzalez, "View-invariant human-body detection with extension to human action recognition using component-wise HMM of body parts," in 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, 2008, pp. 1–6.*
9. *M. N. Kumar and D. Madhavi, "Improved discriminative model for view-invariant human action recognition," Int. J. Comput. Sci. Eng. Technol, vol. 4, no. 3, pp. 1263–1270, 2013.*
10. *T. Syeda-Mahmood, A. Vasilescu, and S. Sethi, "Recognizing action events from multiple viewpoints," in Proceedings IEEE Workshop on Detection and Recognition of Events in Video, 2001, pp. 64–72.*
11. *A. Iosifidis, A. Tefas, and I. Pitas, "Neural representation and learning for multi-view human action recognition," in The 2012 International Joint Conference on Neural Networks (IJCNN), 2012, pp. 1–6.*
12. *R. Poppe, "Vision-based human motion analysis: An overview," Comput. Vis. image Underst., vol. 108, no. 1–2, pp. 4–18, 2007.*
13. *E. Ramasso, C. Panagiotakis, M. Rombaut, D. Pellerin, and G. Tziritas, "Human shape-motion analysis in athletics videos for coarse to fine action/activity recognition using transferable belief model," Electron. Lett. Comput. Vis. Image Anal., vol. 7, no. 4, pp. 32–50, 2009.*
14. *J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997, pp. 928–934.*
15. *Y. Ke, R. Sukthankar, and M. Hebert, "Event detection in crowded videos," in 2007 IEEE 11th international conference on computer vision, 2007, pp. 1–8.*
16. *B. Ni, G. Wang, and P. R. Moulin, "H., 2011. A colour-depth video database for human daily activity recognition," in Proceedings of IEEE International Conference on ComputerVision Workshops, ICCV Workshops, November, pp. 6–13.*
17. *W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in 2010 IEEE computer society conference on computer vision and pattern recognition-workshops, 2010, pp. 9–14.*
18. *C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in 2015 IEEE winter conference on applications of computer vision, 2015, pp. 1092–1099.*

19. *N. E. D. El Madany, Y. He, and L. Guan, "Human action recognition using temporal hierarchical pyramid of depth motion map and keca," in 2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP), 2015, pp. 1–6.*
20. *A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos, "Iberoamerican Congress on Pattern Recognition," 2012.*
21. *O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 716–723.*
22. *S. Lacoste-Julien, F. Sha, and M. I. Jordan, "DiscLDA: discriminative learning for dimensionality reduction and classification. 21st Int Conf on Neural Information Processing Systems," 2008.*
23. *Efros, Berg, Mori, and Malik, "Recognizing action at a distance," in Proceedings Ninth IEEE International Conference on Computer Vision, 2003, pp. 726–733.*
24. *Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 2005, vol. 1, pp. 166–173.*
25. *L. Wang and H. Yu, "Springer Briefs in Molecular Science," Springer Singapore, Singapore, vol. 10, pp. 978–981, 2018.*
26. *J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in 2012 IEEE conference on computer vision and pattern recognition, 2012, pp. 1290–1297.*
27. *R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in 2009 IEEE 12th international conference on computer vision, 2009, pp. 104–111.*
28. *P. Bilinski and F. Bremond, "Contextual statistics of space-time ordered features for human action recognition," in 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance, 2012, pp. 228–233.*
29. *I. Laptev and T. Lindeberg, "Spatial Coherence for Visual Motion Analysis." Springer, 2006.*
30. *S. Zaidenberg, P. Bilinski, and F. Brémond, "Towards unsupervised sudden group movement discovery for video surveillance," in 2014 International Conference on Computer Vision Theory and Applications (VISAPP), 2014, vol. 2, pp. 388–395.*
31. *G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in Workshop on statistical learning in computer vision, ECCV, 2004, vol. 1, no. 1–22, pp. 1–2.*
32. *F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in 2007 IEEE conference on computer vision and pattern recognition, 2007, pp. 1–8.*