INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

# Human Action Recognition Using Dense Trajectories

**Dileep Labana**[1]*, **Kirit Modi**[2]

**1** Ph.D. Student, Computer Science & Engineering, Parul Institute of Technology, Parul University, at & post Vaghodiya, Vadodara, 391760, Gujarat, India
**2** Professor, Computer Engineering Department, Sankalchand Patel University, Visnagar, 384315, Gujarat, India

OPEN ACCESS

**Received:** 21-09-2023

**Accepted:** 03-10-2023

**Published:** 13-11-2023

**Citation:** Labana D, Modi K (2023) Human Action Recognition Using Dense Trajectories. Indian Journal of Science and Technology 16(43): 3846-3853. https://doi.org/10.17485/IJST/v16i43.2408

* **Corresponding author**.

dileep.labana@gmail.com

**Funding:** None

**Competing Interests:** None

Published By Indian Society for Education and Environment (iSee)

**ISSN**
Print: 0974-6846
Electronic: 0974-5645

## Abstract

**Objective:** To develop a robust and effective computer vision system that can automatically identify and classify human actions in video data, considering the temporal dynamics and various environmental conditions. This technology has numerous applications in surveillance, human-computer interaction, and video analysis. **Methods:** The key methods for dense trajectory extraction include the dense optical flow, which computes motion vectors for each point, and the use of key point detectors like the Scale-Invariant Feature Transform (SIFT) or the Harris corner detector. **Findings:** By describing the motion of the trajectories, trajectory descriptors produce remarkably strong results on their own, such as 90.2% on KTH and 47.7% on Hollywood2 for dense trajectories. This demonstrates the significance of the motion data present in the local trajectory patterns. Because the trajectory descriptors catch a lot of camera motion, we only report 67.2% on YouTube. **Novelty:** In this study, a method for modelling movies that combines dense sampling and feature tracking is presented. Compared to earlier video descriptions, our dense trajectories are more reliable. They effectively capture the motion data in the movies and outperform cutting-edge action categorization techniques in terms of performance.

**Keywords:** Human action recognition; Scale-Invariant feature transform; Histograms of oriented gradients; Spatial and temporal interest points; Histograms of optical flow

## 1 Introduction

Computer vision research has seen substantial advancement in the area of human action recognition in videos in recent years, spurred forward by the vast array of practical uses. Video surveillance, video indexing and browsing, gesture detection, human-computer interaction, and sports event analysis are some of these uses. In spite of these developments, the work is still difficult because of a variety of characteristics, such as crowded backdrops, shifting lighting, varied human body types, varying clothes, camera movements, partial occlusions, shifting viewpoints, and size differences in video frames.

Action representation, action learning, and categorization are typically the two main processes in the human action recognition process. According to[1], previous methods of action recognition are based on action representation and may be roughly classified into two groups: global and local representations.

By using backdrop removal or tracking approaches, global representation methods focus on detecting the complete human body. For the localized person, silhouettes, contours, or visual flow are frequently used. These representations are more vulnerable to perspective shifts, changes in physical characteristics, and partial occlusions, however.

Videos are represented using local representation techniques as a collection of tiny, independent patches, each including areas with sizable spatial and temporal changes. Spatial and temporal interest points (STIPs) are the names given to these areas[2]. By extracting appearance and/or motion information from each of the discovered points' corresponding patches, the points are then classified using a dictionary of prototypes or visual words.

The Bag of Words model (BOW) is then used to depict each action sequence[2]. Due to their resounding effectiveness, these strategies have gained a lot of ground in the study of human activity. They successfully solve some of the drawbacks of global representation, including noise sensitivity, partial occlusion, and the requirement for precise localization via background removal and tracking.

## 2 Methodology

It is common practice to use local characteristics to represent videos, especially when paired with a bag-of-features representation, which produces cutting-edge results in action classification. Interest point detectors and local descriptors have recently been effectively used on both pictures and movies.

Videos, however, show different properties in both the 1D temporal domain and the 2D spatial domain. Therefore, treating them differentially makes more sense than employing interest point detection in a single 3D area. Tracking attention spots when watching video sequences offers a simple and effective alternative. Utilising motion data from trajectories, astounding progress in action detection has been shown in recent works[3]. Trajectories are categorized into clusters throughout the video analysis process, and an affine transformation matrix is produced for each cluster center to effectively capture the essence of the trajectories. This was accomplished by extracting trajectories by matching SIFT descriptors between successive frames. They applied a unique-match restriction among the descriptors and eliminated matches that were too far apart from one another to assure correctness.

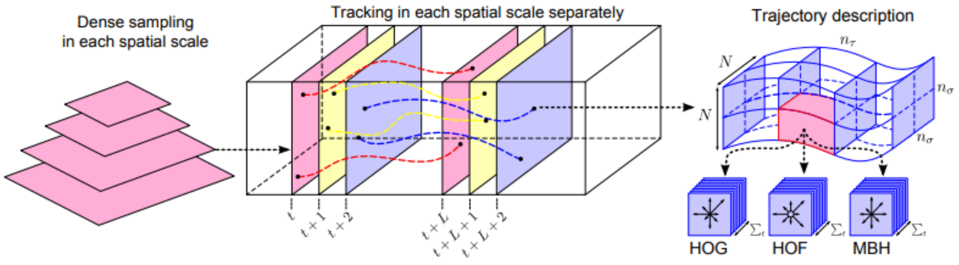

**Fig 1. Example of our dense trajectory description: Left: Feature points are sampled densely for multiple spatial scales; Middle: Tracking is done in the corresponding spatial scale over L frames; Right: Trajectory descriptors are based on its shape represented by relative point coordinates; as well as appearance and motion information over a local neighborhood of NXN pixels along the trajectory**

For picture classification, dense sampling has proven to perform better than sparse interest spots[4]. Similarly, dense sampling at predictable geographical and temporal places fared better than cutting-edge space-time interest point detectors in recent assessments of action identification. However, in order to monitor sparse interest sites, the KLT tracker is commonly used to acquire trajectories. Although successful, matching dense SIFT descriptors is computationally expensive, making it impractical for big video collections.

Our article provides an effective technique for extracting dense trajectories. Tracking highly sampled sites using optical flow fields yields the trajectories. Due to the fact that dense flow fields have already been computed, the number of monitored points may simply be scaled up. In contrast to separately tracking or matching points, Figure 1 shows how the application of global smoothness criteria among points in dense optical flow fields produces more reliable trajectories. Action recognition has not before used dense trajectories. By grouping dense trajectories together, we segmented the objects.

The most useful signal for identifying actions is motion. It could be brought on by the interesting action, but it might also be brought on by the background or camera movement. When dealing with actual behaviors in uncontrolled environments, this is unavoidable. A lingering issue is how to distinguish between motion that is important and motion that is not. The motion correction technique for video stabilization removes the majority of camera motion.

We address the problem of camera motion by providing a local descriptor that emphasizes foreground motion. Our description makes it possible to include dense trajectories into the motion coding method based on motion boundaries that was initially developed for human detection. We demonstrate that motion bounds recorded along the trajectories perform noticeably better than cutting-edge descriptors. The structure of this essay is as follows: We introduce the method for obtaining dense trajectories in Section 2. Then, in section 3, we demonstrate how to encode feature descriptors throughout the trajectory. Finally, in sections 3 and 4, respectively, we provide the experimental design and talk about the findings.

## 2.1 Dense Trajectories

Figure 1 shows the extraction of dense trajectories at various spatial scales. Feature points are sampled and each scale tracks them independently on a grid with W-pixel spacing. Through investigation, we discovered that a sample step size of W = 5 is adequate to produce reliable results. We used 8 spatial scales that were divided in half. Every point Pt = (Xt, Yt) at frame t is tracked to the frame t+ 1 after median filtering in a dense optical flow field = (ut), vt).

$$Pt+1 = (Xt+1,\ Yt+1) = (Xt+Yt) + (M*\omega)|\ (Xt,\ Yt) \tag{1}$$

A median filtering kernel M is used to round (Xt, Yt) in order to obtain the position (Xt, Yt). In particular, for sites close to motion boundaries, this method is more reliable than the bilinear interpolation method used in Points may be monitored in great detail once the dense optical flow field has been generated without adding more computing work. Points from the following frames are combined to form a trajectory: (Pt, Pt+1, Pt+2). We employ the Farneback technique as implemented in the OpenCV library2 to extract dense optical flow. This method, in our opinion, strikes a decent balance between accuracy and speed.

Trajectories that stray from their starting places are a challenge that tracking frequently faces. We handle the length of each trajectory by restricting it to L frames in order to alleviate this problem. As seen in Figure 2 (middle), once a trajectory reaches this length, it is removed from the tracking process. We check each frame for the existence of a track on our thick grid to ensure complete coverage across the video. If a W-W neighborhood has no monitored points, a feature point is sampled and used in the tracking procedure. We used a trajectory length of L = 15 frames for our experiment.

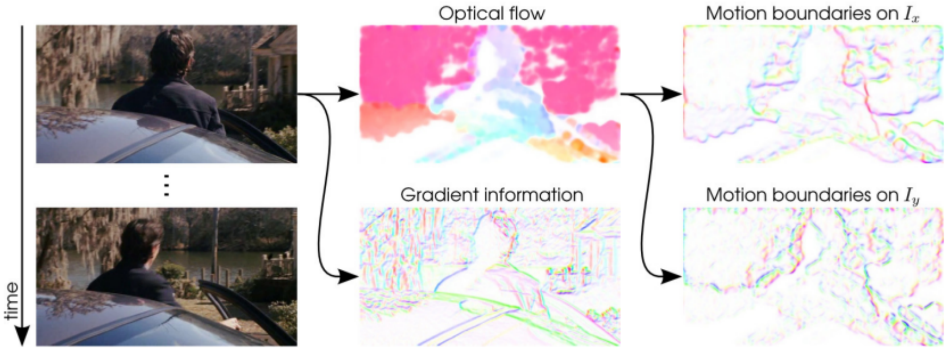

**Fig 2. Example of the data gathered by the HOG, HOF, and MBH descriptors. Color (hue) and saturation serve as indicators of the gradient/flow direction and magnitude for each picture. Motion boundaries are calculated as the individual x and y optical flow gradients. Motion boundaries, as opposed to optical flow, emphasize foreground motion while suppressing the majority of background camera motion. Motion bounds, as opposed to gradient information, remove the majority of texture data from the static backdrop**

It is impossible to follow points in areas of uniform images devoid of any pattern[5]. We examine the lower eigenvalue of an autocorrelation matrix after sampling a feature point. Additionally, erroneous trajectory with sudden and substantial displacements is deleted[6]. The dense and KLT trajectories are compared in Figure 1. We can see that dense trajectories outperform those produced by the KLT tracker in terms of robustness and density.

Local motion patterns are encoded in a trajectory's form. Given a trajectory of length L, we may determine its form from a sequence of displacement vectors Pt = (Pt+1 Pt) = (xt+1 xt, yt+1 yt) called S = (Pt, Pt+L1). The sum of the displacement vectors' magnitudes is used to normalize the resultant vector:

$$S' = \frac{(\triangle Pt, \ldots\ldots\ldots\ldots \triangle Pt + L - 1)}{\sum_{j=1}^{t+L-1} ((\triangle Pj||} \tag{2}$$

This vector is referred to as a trajectory description. In order to distinguish between activities that are carried out at various speeds, we have also considered expressing trajectories at various temporal scales. In reality, this had no positive impact on the outcomes. As a result, in our studies, we adopt trajectory types with fixed length L.

### 2.2 Descriptors with a trajectory

As a method of representing video, local descriptors calculated inside a 3D video volume near interest spots have grown in favor[7]. Figure 2 illustrates how we compute descriptors inside a space-time container enclosing the trajectory in order to maximize the motion information in our dense trajectories. This volume spans L frames and is NXN pixels in size. We split the volume into a spatio-temporal grid of dimension nxn in order to add structural information to the representation. We employ the default values N = 32, n = 2, and n = 3 for our studies because cross-validation on the Hollywood2 training set has shown them to be the most effective.

Among the well-known descriptors, HOGHOF has demonstrated itself to be extremely successful for action recognition across a variety of datasets. HOG (histograms of oriented gradients)[8] concentrates on the static appearance features, whereas HOF (histograms of optical flow) primarily collects local motion information. We calculate the HOGHOF description along with our dense trajectories. An additional ninth bin is utilized for HOF, making a total of nine bins. Orientations are quantized for HOG and HOF into 8 bins that include full orientations. The L2 norm is used to normalize both descriptors. A depiction of the HOGHOF is shown in Figure 2.

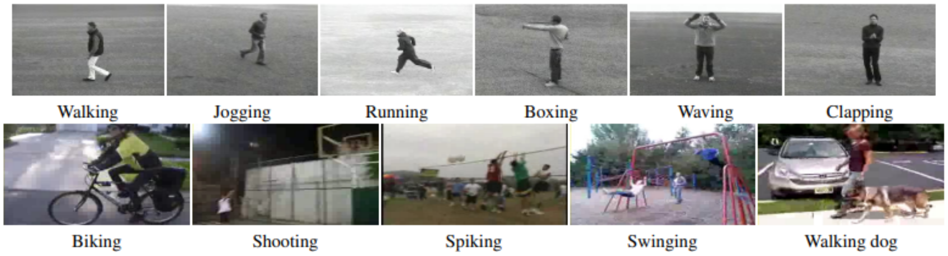

**Fig 3. Examples of video frames from action datasets from KTH (top row), YouTube (second row), Hollywood2 (third row), and UCF sports (last row)**

The MBH descriptor separates the optical flow field I = (Ix, Iy) into its x and y components. The orientation data is quantized into histograms, and, like the HOG descriptor, spatial derivatives are computed for each of them. We obtain an 8-bin histogram for each component, and we separately normalize each one using the L2 norm. Because MBH reflects the gradient of the optical flow and suppresses information about steady motion, it only retains information about changes in the flow field, such as motion boundaries. Compared to video stabilization[9] and motion correction, this is a simple way to remove noise caused by background motion.

## 3 Results and Discussion

We first discuss the datasets utilized for action recognition in this section. The bag-of-features methodology used to assess our dense trajectory characteristics and the KTL tracking baseline is then briefly presented.

### 3.1 Dataset

Figure 3 shows an exhaustive evaluation of our dense trajectories on four common action datasets: KTH, YouTube, Hollywood2, and UCF sports. The datasets here are highly varied. While the Hollywood2 dataset includes actual movies with a cluttered background, the KTH dataset shows activities against a homogeneous background. The UCF sports videos are excellent resolution, in contrast to the low-quality YouTube videos.

Six human activity classes—walking, jogging, sprinting, boxing, waving, and clapping—are included in the KTH dataset. 25 individuals execute each activity many times. The four circumstances in which the clips were captured were outdoors, outdoors with scale variation, outdoors with various attire, and inside. Most segments have uniform and motionless backgrounds[10]. The data consists of 2391 video clips in total. Following the authors' initial experimental design, we split the data into a training set (the remaining 16 participants) and a test set (9 subjects: 2, 3, 5, 6, 7, 8, 9, 10, and 22). We train and assess a multi-class classifier, the same as we did in the first study, and we present average accuracy across all classes as a performance metric.

11 different activity genres are included in the YouTube dataset: basketball shooting, biking/cycling, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline leaping, volleyball spiking, and walking a dog. Large changes in camera movements, item size, shape, and attitude, perspective, a crowded background, and lighting conditions make this dataset tough. There are 1168 sequences altogether in the collection. We use leave-one-out cross-validation for a predetermined set of 25 folds in accordance with the original configuration. As a performance indicator, the overall average accuracy for all courses is presented.

A total of 69 distinct Hollywood films made up the Hollywood2 dataset[11]. There are 12 action classes: picking up the phone, operating a vehicle, eating, engaging in conflict, exiting a vehicle, shaking hands, hugging, and kissing, as well as moving, sitting, sitting up, standing, and sprinting. We utilized the clean training dataset for our research. A training set (823 sequences) and a test set (884 sequences) comprise the entire 1707 action samples. Different movies are used for the train and test segments. By calculating the average accuracy (AP) for each action class and providing the mean AP across all classes (MAP), the performance is assessed.

Ten human movements are included in the UCF sport dataset: swinging (on the pommel horse and on the ground), diving, kicking (a ball), lifting weights, riding a horse, running, skateboarding, swinging (at the high bar), swinging a golf club, and walking. 150 video clips make up the dataset, which exhibits significant intra-class heterogeneity. We expand the dataset by including a horizontally flipped version of each sequence in order to increase the number of data samples. We train a multi-class classifier using data from the KTH action and then present the average accuracy across all classes. We employ a leave-one-out arrangement, testing each original sequence while training each other's flipped versions of all other sequences (i.e., the flipped version of the tested sequence is excluded from the training process).

### 3.2 Collection of attributes

We employ the common collection of attributes method to assess how well our dense trajectory's function for each descriptor (trajectory, HOG, HOF, and MBH) independently[12], we first create a codebook. We chose the maximum number of visual words per descriptor at 4000 since this amount has been empirically proven to produce satisfactory results for a variety of datasets. We use k-means to cluster a subset of 100,000 randomly chosen training characteristics in order to reduce the complexity. We run k-means eight times and keep the result with the lowest error to enhance precision. Euclidean distance is used to assign descriptors to the nearest vocabulary term. As video descriptors, the generated histograms of visual word occurrences are employed.

Multichannel technique is used to mix several descriptors, as in[13]:

$$K\left(x_i, y_j\right) = exp(-\ \sum_c \frac{1}{A^c} D\left(x_i^c, y_j^c\right)) \tag{3}$$

where $D\left(x_i^c, y_j^c\right)$ is the c-th channel-dependent 2 distances between video xi and xj. $A^c$ is the average value of the c-th channel's training samples' distances of 2. When dealing with several classes, we employ a one-against-rest strategy and choose the class with the greatest score.

### 3.3 Results from experiment

In this part, we assess the effectiveness of our description and compare it to cutting-edge techniques. We also assess the impact of various parameter values.

### 3.4 An assessment of our dense trajectory descriptors

In this part, we compare the various descriptors and the dense and KLT trajectories. In order to do this comparison, we utilize our default settings. We used N = 32, n = 2, and n = 3 for baseline KLT and dense trajectories to construct the descriptors. We set the dense sampling step size to be W = 5 and the trajectory length to be L = 15. Table 1 displays the outcomes for the four datasets. Overall, our dense trajectories perform 2% to 6% better than KLT trajectories. This shows that our dense trajectories represent the video structures more precisely since the descriptors are the same.

Only describing the motion of the trajectories, trajectory descriptors produce remarkably strong results on their own, such as 90.2% on KTH and 47.7% on Hollywood2 for dense trajectories. This demonstrates the significance of the motion data present in the local trajectory patterns. Because the trajectory descriptors catch a lot of camera motion, we only report 67.2% on YouTube. In general, HOF performs better than HOG because motion is more discriminative for recognizing actions than static appearance. HOG, though, does well on UCF sports and YouTube. Since many YouTube videos are captured with hand-held cameras, the HOF descriptors calculated on those films are highly contaminated by camera movements. For UCF sports activities, which frequently entail specialized equipment and scene kinds, static scene context is crucial. On all four datasets, MBH consistently outperforms the other descriptors. The uncontrolled realistic datasets YouTube and Hollywood2 show the biggest progress. On YouTube, for instance, MBH is 11.1% superior to HOF. This demonstrates the benefit of reducing background motion while handling optical flow.

**Table 1. KLT and dense trajectory comparisons, as well as comparisons of other descriptions on KTH, YouTube, Hollywood2, and UCF sports**

| | **KTH** | | **YouTube** | | **Hollywood** | | **UCF Sports** | |
|---|---|---|---|---|---|---|---|---|
| | **KLT** | **Dense Trajectories** | **KLT** | **Dense Trajectories** | **KLT** | **Dense Trajectories** | **KLT** | **Dense Trajectories** |
| **TRAJE.** | 88.40% | 90.20% | 58.20% | 67.20% | 46.20% | 47.70% | 72.80% | 75.20% |
| **HOG** | 84.00% | 86.50% | 71.00% | 74.50% | 41.00% | 41.50% | 80.20% | **83.80%** |
| **HOF** | 92.40% | 93.20% | 64.10% | 72.80% | 48.40% | 50.80% | 72.70% | 77.60% |
| **MBH** | 93.40% | **95.00%** | 72.90% | **83.90%** | 48.60% | **54.20%** | 78.40% | **84.80%** |
| **Our Approach** | **93.20%** | **94.10%** | **79.60%** | **84.10%** | **54.50%** | **58.20%** | 82.00% | **88.10%** |

**Table 2. Accuracy for the YouTube dataset per action class.**

| | **KLT** | **Dense trajectories** | **Ikizler-Cinbis** [9] |
|---|---|---|---|
| B_shoot | 33.0% | 42.0% | **47.48%** |
| Bike | 86.6% | **90.7%** | 74.17% |
| Dive | **98.0%** | **98.0%** | 95.0% |
| Golf | 96.0% | **97.0%** | 96.0% |
| Hride | 77.0% | **85.0%** | 72.0% |
| Sjuggle | 64.0% | **76.0%** | 53.0% |
| Swing | 85.0% | **88.0%** | 66.0% |
| Tswing | 70.0% | 71.0% | **77.0%** |
| Tjump | 92.0% | **94.0%** | 92.0% |
| Vspike | **96.0%** | 95.0% | 84.0% |
| Walk | 75.4% | **87.0%** | 65.67% |
| **Accuracy** | 78.9% | **83.2%** | 74.21% |

**Table 3. Per action class average accuracy for the Hollywood dataset**

| | KLT | Dense trajectories | Ullah [13] |
|---|---|---|---|
| Answerphone | 18.3% | **32.6%** | 25.9% |
| Drive Car | **88.8%** | 88.0% | 85.9% |
| Eating | **73.4%** | 65.2% | 56.4% |
| Fight Person | 74.2% | **81.4%** | 74.9% |
| GetOut_Car | 47.9% | **52.7%** | 44.0% |
| Hand_Shake | 18.4% | 29.6% | **29.7%** |
| Hug Person | 42.6% | **54.2%** | 46.1% |
| Run | 76.3% | **82.1%** | 69.4% |
| Sit_down | 59.0% | **62.5%** | 58.9% |
| Sit_Up | **27.7%** | 20.0% | 18.4% |
| Stand_Up | 63.4% | **65.2%** | 57.4% |
| MAP | 54.5% | **58.2%** | 51.8% |

## 3.5 Comparison to the state-of-the-art

We contrast the outcomes for each action class on YouTube. When compared to the KLT baseline and the technique of[13–15], our dense trajectories on YouTube produce the best results for 8 out of the 11 action classes, as shown in Table 2. See Table 3 for a comparison of the AP for each action class on Hollywood2 with the KLT baseline and the technique, which combines 24 spatiotemporal grids. The best outcomes come from our dense trajectories for 8 out of the 12 action classes.

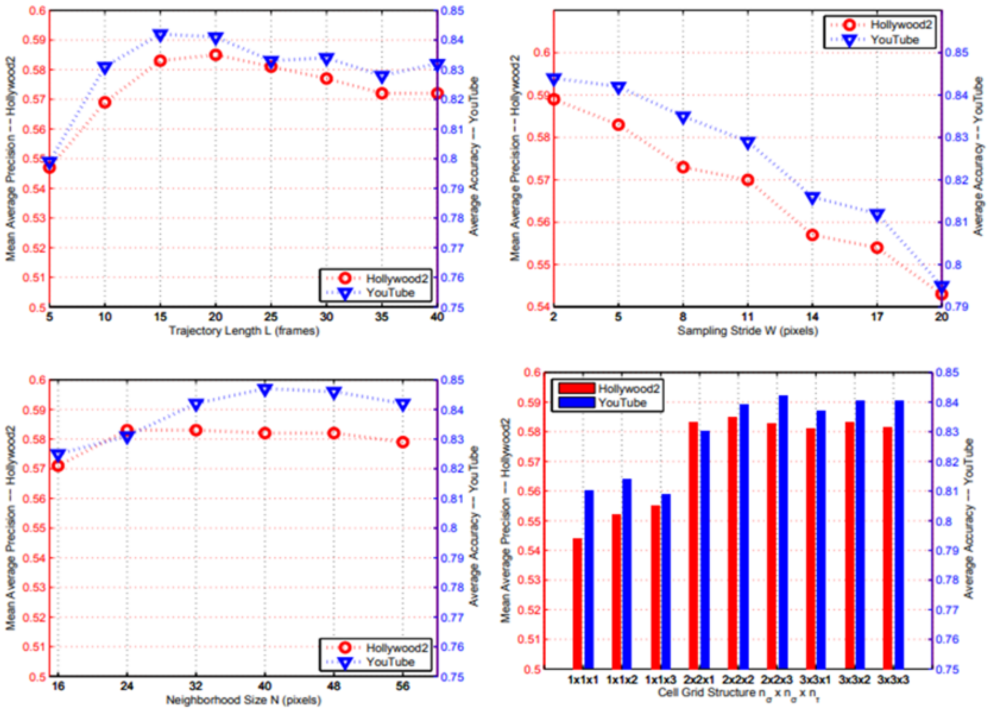

**Fig 4. Results on the Hollywood and YouTube datasets with various parameter values**

## 4 Conclusion

In this study, a method for modelling movies that combines dense sampling and feature tracking is presented. Compared to earlier video descriptions, our dense trajectories are more reliable. They effectively capture the motion data in the movies and outperform cutting-edge action categorization techniques in terms of performance. By generating motion boundary descriptors along the dense trajectories, we have also devised an effective way to eliminate camera motion. Our trajectory descriptors produce remarkably strong results on their own, such as 90.2% on KTH and 47.7% on Hollywood2 for dense trajectories. This demonstrates the significance of the motion data present in the local trajectory patterns. Because the trajectory descriptors catch a lot of camera motion, we only report 67.2% on YouTube. This beats earlier video stabilizing techniques and effectively separates relevant motion from background motion. Our descriptors incorporate information on trajectory shape, appearance, and velocity. A similar form has proven effective for classifying actions, but it might also be used for other tasks like action localization and video retrieval.Future research should put a focus on integrating deep learning methods, improving temporal modelling, addressing ethical issues, and expanding recognition capabilities to include 3D trajectories, interactions, and group behaviors.

## References

1) Agarwal S, Gupta MK. Context Aware Image Sentiment Classification using Deep Learning Techniques. *Indian Journal Of Science And Technology*. 2022;15(47):2619–2627. Available from: https://doi.org/10.17485/IJST/v15i47.1907.
2) Patel CI, Labana D, Pandya S, Modi K, Ghayvat H, Awais M. Histogram of Oriented Gradient-Based Fusion of Features for Human Action Recognition in Action Video Sequences. *Sensors*. 2020;20(24):1–32. Available from: https://doi.org/10.3390/s20247299.
3) Srilakshmi N, Radha N. An Enhancement of Deep Positional Attention-Based Human Action Recognition by Using Geometric Positional Features. *Indian Journal Of Science And Technology*. 2023;16(29):2190–2197. Available from: https://doi.org/10.17485/IJST/v16i29.379.
4) Xu Y, Zhou F, Wang L, Peng W, Zhang K. Optimization of Action Recognition Model Based on Multi-Task Learning and Boundary Gradient. *Electronics*. 2021;10(19):1–16. Available from: https://doi.org/10.3390/electronics10192380.
5) Patel R, Vaghela R, Chopade M, Patel P, Bhatt D. Integrated Neuroinformatics: Analytics and Application. In: Knowledge Modelling and Big Data Analytics in Healthcare. Boca Raton. CRC Press. 2021. Available from: https://www.taylorfrancis.com/chapters/edit/10.1201/9781003142751-9/.
6) Labana D, Modi K. Human Action Recognition via Multi-Task Learning. *Journal of Emerging Technologies and Innovative Research*. 2023;10(7):409–414. Available from: https://www.jetir.org/papers/JETIR2307548.pdf.
7) Papadopoulos K, Demisse G, Ghorbel E, Antunes M, Aouada D, Ottersten B. Localized Trajectories for 2D and 3D Action Recognition. *Sensors*. 2019;19(16):1–22. Available from: https://doi.org/10.3390/s19163503.
8) Nguyen TT, Nguyen TP, Bouchara F. Directional dense-trajectory-based patterns for dynamic texture recognition. *IET Computer Vision*. 2020;p. 162–176. Available from: https://doi.org/10.1049/iet-cvi.2019.0455.
9) Arif S, Ul-Hassan T, Hussain F, Wang J, Fei Z. Video representation by dense trajectories motion map applied to human activity recognition. *International Journal of Computers and Applications*. 2020;42(5):474–484. Available from: https://doi.org/10.1080/1206212X.2018.1486001.
10) Morceli BDM, Poz APD. Road extraction from low-cost GNSS-device dense trajectories. *Journal of Location Based Services*. 2023;17(3):251–270. Available from: https://doi.org/10.1080/17489725.2023.2216670.
11) Camarena F, Chang L, Gonzalez-Mendoza M, Cuevas-Ascencio RJ. Action recognition by key trajectories. *Pattern Analysis and Applications*. 2022;25(2):409–423. Available from: https://link.springer.com/content/pdf/10.1007/s10044-021-01054-z.pdf?pdf=button.
12) Yi Y, Li A, Zhou X. Human action recognition based on action relevance weighted encoding. *Signal Processing: Image Communication*. 2020;80:115640. Available from: https://doi.org/10.1016/j.image.2019.115640.
13) Zhao H, Dang J, Wang S, Wang Y, Gao D. Dense Trajectory Action Recognition Algorithm Based on Improved SURF. *IOP Conference Series: Earth and Environmental Science*. 2019;252(3):1–8. Available from: https://iopscience.iop.org/article/10.1088/1755-1315/252/3/032179/meta.
14) Roseline V, Chellam GH. A Novel Fusion Attention Algorithm for Sentimental Image Analysis. *Indian Journal of Science and Technology*. 2022;15(9):386–394. Available from: https://doi.org/10.17485/IJST/v15i9.2159.
15) Mefteh S, Kaâniche MBB, Ksantini R, Bouhoula A. A novel multispectral corner detector and a new local descriptor: an application to human posture recognition. *Multimedia Tools and Applications*. 2023;82:28937–28956. Available from: https://doi.org/10.1007/s11042-023-14788-1.