electronics

Article

Machine Vision-Based Human Action Recognition Using Spatio-Temporal Motion Features (STMF) with Difference Intensity Distance Group Pattern (DIDGP)

Jawaharlalnehru Arunnehru [1,*], Sambandham Thalapathiraj [2], Ravikumar Dhanasekar [3], Loganathan Vijayaraja [3], Raju Kannadasan [4], Arfat Ahmad Khan [5,*], Mohd Anul Haq [6,*], Mohammed Alshehri [7], Mohamed Ibrahim Alwanain [8] and Ismail Keshta [9]

1 Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani Campus, Chennai 600026, India
2 Department of Mathematics, SRM Institute of Science and Technology, Vadapalani Campus, Chennai 600026, India; thalapathirajs@gmail.com
3 Department of Electrical and Electronics Engineering, Sri Sairam Institute of Technology, West Tambaram, Chennai 600044, India; dhanasekar.eee@sairamit.edu.in (R.D.); vijayaraja.eee@sairamit.edu.in (L.V.)
4 Department of Electrical and Electronics Engineering, Sri Venkateswara College of Engineering, Chennai 602117, India; kannadasanr@svce.ac.in
5 College of Computing, Khon Kaen University, Khon Kaen 40000, Thailand
6 Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Al-Majmaah 11952, Saudi Arabia
7 Department of Information Technology, College of Computer and Information Sciences, Majmaah University, Al-Majmaah 11952, Saudi Arabia; ma.alshehri@mu.edu.sa
8 Department of Computer Science, College of Science and Humanities in Alghat, Majmaah University, Al-Majmaah 11952, Saudi Arabia; m.alwanin@mu.edu.sa
9 Computer Science and Information Systems Department, College of Applied Sciences, AlMaarefa University, Ad Diriyah, Riyadh 13713, Saudi Arabia; imohamed@mcst.edu.sa
* Correspondence: arunnehru.aucse@gmail.com (J.A.); arfatkhan@kku.ac.th (A.A.K.); m.anul@mu.edu.sa (M.A.H.)

Abstract: In recent years, human action recognition is modeled as a spatial-temporal video volume. Such aspects have recently expanded greatly due to their explosively evolving real-world uses, such as visual surveillance, autonomous driving, and entertainment. Specifically, the spatio-temporal interest points (STIPs) approach has been widely and efficiently used in action representation for recognition. In this work, a novel approach based on the STIPs is proposed for action descriptors i.e., Two Dimensional-Difference Intensity Distance Group Pattern (2D-DIDGP) and Three Dimensional-Difference Intensity Distance Group Pattern (3D-DIDGP) for representing and recognizing the human actions in video sequences. Initially, this approach captures the local motion in a video that is invariant to size and shape changes. This approach extends further to build unique and discriminative feature description methods to enhance the action recognition rate. The transformation methods, such as DCT (Discrete cosine transform), DWT (Discrete wavelet transforms), and hybrid DWT+DCT, are utilized. The proposed approach is validated on the UT-Interaction dataset that has been extensively studied by past researchers. Then, the classification methods, such as Support Vector Machines (SVM) and Random Forest (RF) classifiers, are exploited. From the observed results, it is perceived that the proposed descriptors especially the DIDGP based descriptor yield promising results on action recognition. Notably, the 3D-DIDGP outperforms the state-of-the-art algorithm predominantly.

Keywords: feature extraction; human activity recognition; spatio-temporal features; support vector machines; video surveillance

Citation: Arunnehru, J.; Thalapathiraj, S.; Dhanasekar, R.; Vijayaraja, L.; Kannadasan, R.; Khan, A.A.; Haq, M.A.; Alshehri, M.; Alwanain, M.I.; Keshta, I. Machine Vision-Based Human Action Recognition Using Spatio-Temporal Motion Features (STMF) with Difference Intensity Distance Group Pattern (DIDGP). *Electronics* **2022**, *11*, 2363. https://doi.org/10.3390/electronics11152363

Academic Editor: George A. Papakostas

Received: 19 June 2022
Accepted: 18 July 2022
Published: 28 July 2022

1. Introduction

Video acquisition technologies are becoming pervasive in our daily lives. Powerful digital cameras used in social media, traffic, security, and emergency monitoring are capable of capturing high-level details of people's faces and body posture for activity recognition [1]. However, these approaches model machine learning architectures requiring high computational power, disregarding real-time performance and integration into embedded devices. The primary motivation of Human action recognition is extensively investigated in real-time, and it represents a vital role in many applications such as video surveillance, human-computer interaction, and video content retrieval [2]. Human motion, typically a combination of interpretation and rotational motions of each body joint contains much information inherent to humans [3,4]. In selective, motion similarity can be accomplished by analyzing the human motions with various applications. For example, motion similarity can be used for action recognition, and it is also possible to measure a motion similarity to conclude whether a task is performed well or to identify abnormal behavior [5]. A motion estimation system helps to match a target person from different cameras for re-identification. While interpreting, human motion plays an imperative role in the tasks mentioned, motion similarity research has attracted less attention so far due to the subsequent reasons. Firstly, measuring the motion similarity is a challenging problem. Different camera views or human body structures cause various 2D joint coordinates even for similar motions in videos. This makes it difficult to measure the similarity using the joint coordinates directly. Secondly, the availability of large-scale datasets for learning the motion similarity is limited. Lastly, there are few human motion datasets available for assessing the performance of different motion similarities to computation methods. Compared with traditional images, spatial temporal-based action recognition has drawn increasingly more attention since it is robust against embellished backgrounds and camera viewpoints.

Human body motions are represented as a sequence of 2D or 3D spatial coordinates, and they provide a good representation for describing human actions with motion or texture features [6]. The spatial data can be easily obtained by video cameras or pose estimation algorithms. The motion information of the joints is also an important cue to recognize the underlay action. Some actions such as "hugging" and "punching" are challenging to recognize from spatial information and can extort the movements of body joints to help the recognition. Since the spatial data are represented as the action coordinates, joints motion is easily calculated as the difference of coordinates along the temporal dimension Figure 1 shows the three categories of human motion analysis. Hence, this work establishes transform-based approaches for human action recognition. Several studies were reported on feature descriptor, representation, and classification methods for robust action recognition [7–11]. Knowing human behaviors remains a challenge due to diverse complex variables, such as perspective, size, rotation, shifts in tempo, different differences in the anthropometry, and embroiled contexts. The human body usually varies dramatically in size, physique, and appearance in different groups of actors while performing the same action.

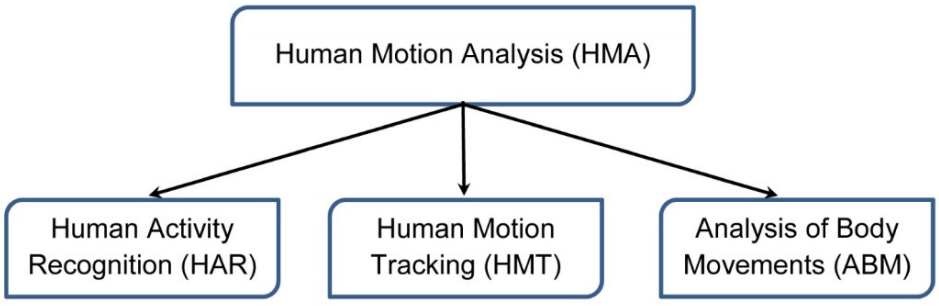


Figure 1. Categories of human motion analysis.

The biggest challenge in recognizing human activity is identifying and extracting the right and significant features. The Deep learning technique that has recently been

created can extract, as well as choose, the relevant features. The convolutional neural network (CNN) is one of the many deep learning techniques that have the benefits of local dependency and scale invariance and are appropriate for temporal data. The CNN is excelling at handling the temporal data when compared to the traditional machine learning techniques, which demands on domain specific knowledge [12,13].

2. Related Work

Several studies [1–3,14,15] has been conducted in recent years in the field of recognition of human behavior to reduce the manual effort and increase computational performance. Laptev [16] and Dollar [17] proposed a space-time interest point detector for action recognition and these feature points showed discriminative properties like appearance and positions. Feature extraction methods for identifying behavior can generally be divided into four categories: geometry based [18,19], motion-based [20,21], appearance-based [22,23], and space-time feature-based [17,24–26]. The geometry-based approaches use geometric points from human body structure that is normally difficult and time-consuming for object segmentation and tracking. The movement-based optical flow models for process recognition, but it reduces the effects of background flows due to foreground segmentation. Motion patterns have also been considered as an important cue for action recognition. The appearance-based approaches use silhouette information to recognize actions, but they are weak to cluttered backgrounds. The space-time feature-based approaches uses space-time interest points to distinguish action categories. The majority of the work in human action recognition has been conducted on the standard benchmark dataset, such as the KTH dataset [24] and Weizmann dataset [27], which consists of various kinematics activities performed by a single actor with homogeneous backgrounds with some datasets like HMDB51 [28] and Hollywood2 [29] that comprises of realistic environments that have been used for evaluation. Furthermore, another type of benchmark dataset such as UT-Interaction [30] is a human-human interaction activity.

Laptev [16] used the idea of Harris and Forstner's point-of-interest using operators to identify the local structure with significant space and time-domain deviations. Dollar [17] used linear separable filters for detecting interest points in the local region, which respond to strong motion and the space-time corners. Moreover, it was proposed to use a Hough transform-based voting framework for action recognition that used Spatio-temporal voting with extracted local $X - Y - T$ features [31]. They performed recognition by voting with a collection of random learned trees in Hough space. In some works [32], the SVM classification method with χ^2 kernels proposed a structure on dense multiscale trajectories that extracted the dense trajectory (DT) form, gradient orientation histogram (HoG), flow orientation histogram (HoF), and motion boundary histogram (MBH) from image data. Finally, the visual Code-book was read from the training models. There were two well-known standard methods in the frequency-domain, discrete cosine transformation (DCT) and discrete wavelet transformation (DWT). The DCT [33] technique had been extensively used in the field of digital image processing field that involved compression of images including various enhancement techniques and segmentation. It is similar to discrete fourier transform (DFT) that concentrated the massive DCT coefficients into the low-frequency region and had excellent characteristics of energy properties. DWT [34] used an orthogonal rule that can be applied to separate finite data into different frequency components such as approximation coefficient matrix (cA) and other coefficients matrices are horizontal coefficient (cH), vertical coefficient (cV), and diagonal coefficient (cD).

Some of the work [35] presented a method based on mean and variance that was computed from the absolute DCT coefficients values for the entire image using texture-based classification. It also proposed a content-based image retrieval method based on quad tree-structure using DCT coefficients as quad tree nodes to represent the image features [36]. Here, the wavelets helped in detecting significant points by representing the local properties of images. It used the multi-resolution wavelet decomposition that extracted the intensity-hue-saturation (IHS) and principal components analysis (PCA) to implement the

spatial detail regarding wavelet-based image fusion [37]. The transformation methods eliminated redundancy in neighboring pixels that offered the advantage of determining the uncorrelated transform coefficients. Hence, the main advantage of this process is to reduce the correlation between neighboring pixels which in turn leads to irrelevant coefficients of transformation. The existing works highlighted that the geometry-based approaches were time-consuming for object segmentation and tracking the spatiotemporal interest points in three-dimensional patterns. Previous researchers limited their work to 3–4 actions since many databases do not prefer a wide variety of action sequences. Previous human activity recognition approaches lack pre-processing steps that effectively filter data and increase the classifier accuracy. Hence, this paper proposes a novel approach based on the STIPs for action descriptors for representing and recognizing human actions in video sequences.

2.1. Contributions

This work discusses the action recognition problem by extracting spatio-temporal interest points. Initially, the approach captures the local motion in a video that is invariant to size and shape changes. Then, classification methods, such as Support Vector Machines (SVM) and Random Forest (RF) classifiers, are exploited. The approach extends further to build unique and discriminative feature description methods along with PCA for feature dimensional reduction in order to enhance the action recognition rate. In this work, a novel approach based on the STIPs is proposed for action descriptors i.e., Two Dimensional-Difference Intensity Distance Group Pattern (2D-DIDGP) and Three Dimensional-Difference Intensity Distance Group Pattern (3D-DIDGP) for representing and recognizing the human actions in video sequences. In addition to that, the transformation methods such as DCT, DWT, and hybrid DWT+DCT are utilized. Predominantly, the 3D-DIDGP method outclasses the state-of-the-art algorithm.

2.2. Organization

The article begins with background, literature report, and contributions relating to the proposed work. Following this, the proposed method is described in detail in Section 3. Further, this work presents a detailed experimental setup for the study in Section 4. Then, Section 5 illustrates the test results under different methods of operation. Subsequently, a comparative analysis from the observed results is carried out in Section 6. Finally, Section 7 concludes the article with the key observations.

3. Proposed Method

Motion patterns in the field of action recognition were identified based on changes in a subject's location regarding time. Since motion information is an important cue to describe the action. Initially the input video is converted to gray scale and further the noise is removed for fine features. All frames are smoothed by Gaussian convolution method with a matrix size of 5×5 for successful feature extraction and classification. The frame difference method is adapted to extract the motion features. In this work, difference intensity distance group pattern (DIDGP) based 2D/3D cuboids extraction and transformation based DCT, DWT, Hybrid DWT+DCT also applied at each spatial-temporal interest point of an action sequence. Moreover, principal component analysis (PCA) is adapted to select the most discriminatory motion features to improve action recognition performance. Finally, a Support Vector Machine (SVM) and Random Forest classifier are used to classify the actions.

3.1. Identifying Motion by Frame Differencing

In video analysis research, object detection still remains as open problem. Since the objects in a video are moving in general, if the object moves from the camera viewpoint, the images of the object may dramatically differ. This change may arise due to variation in target pose, variation in illumination and partial and total occlusion of the target.

Initially, the input video is converted into frames and the extracted color frames are converted to grey-scale frame using simple average method by adding the pixel values of red (R), green (G) and blue (B) channels and divided by three: (R + G + B)/3. Secondly, the 2D Gaussian smoothing operator is used to 'smooth' images and remove detail and noise. In this sense it is similar to the mean filter, with help of 5 × 5 kernel. This kernel has some special properties. Third, Frame differencing is defined by the difference between consecutive frames in time, instead of subtracting predefined or estimated background on-the fly, the frame subtraction method considers every pair of frames in time t and $t + 1$, and extracts any motion in it. In order to find any region of interest/object present in a video frame, simply subtract the current frame with the previous frame on pixel-by-pixel basis. The difference image computed between the two consecutive frames of the video. Difference image at time t is given by

$$D_t(x, y) = |I_t(x, y) - I_{t+1}(x, y)| \quad 1 \leq i \leq w,\ 1 \leq j \leq h \tag{1}$$

3.2. Overview of the Proposed Human Action Recognition Framework

Figure 2 shows the description of the method proposed. The approach begins by extracting the motion information followed by identifying the interest points from the training video sequences. Depending on the structural distribution of interest points, each sequence generates the descriptors in the cuboid. Moreover, another set of features called DIDGP in 2D and 3D are extracted from the training sequences. The above measures are repeated during the process. For classifying the test sequences to the suitable kind of behavior based on the model developed in the training stage, SVM and Random Forest Classifiers are adopted.

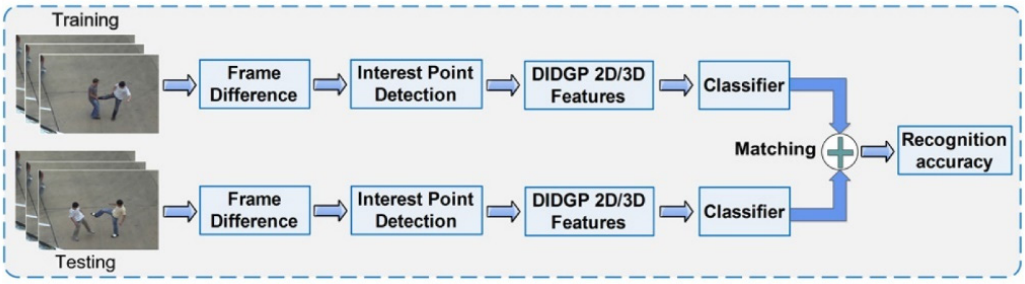


Figure 2. Overview of the proposed method.

Interest Point Identification

This work applies the Harris interest point detector [38] due to its strong in-variance to rotation, scaling, illumination, and noise. Interest points in a video are constrained along both spatial and temporal dimensions.

Harris corner detector is based on the local auto-correlation function. At a corner, the image intensity will change largely in multiple directions. For the image I, the algorithm calculates the change of intensity for the shift $[u, v]$ as follows:

$$E(u, v) = \sum_{x,y} w(x, y)[I(x + u,\ y + v) - I(x, y)]^2 \tag{2}$$

where $w(x, y)$ is called as window at $(x, y)I(x, y)$ is the intensity at $(x, y)I(x + u,\ y + v)$ is the intensity of moved window $(x + u,\ y + v)$.It is required to capture corners with a maximum variation in intensity. Hence, the shifted image is approximated by a Taylor expansion and finally a score is calculated to determine interesting point as represented in the following equation,

$$R = \det(M) - k(trace(M))^2 \tag{3}$$

where $\det(M) = \lambda_1\lambda_2 trace(M) = \lambda_1 + \lambda_2$. A window with score R greater than a threshold is considered as 'interest point'. Figure 3 shows the interest points detected in 'hug' sequence in the UT-Interaction dataset. The highlighted points relating to the local maxima response function known as spatio-temporal points of interest. Cuboids obtained from the various actions are shown in Figure 4. It is also evident from Figure 5 that the actions can be clearly distinguished from these cuboids.

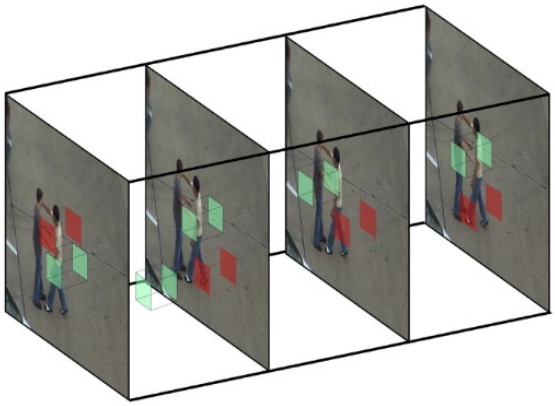


Figure 3. The view of cuboids extracted around each spatial temporal interest point for each action.

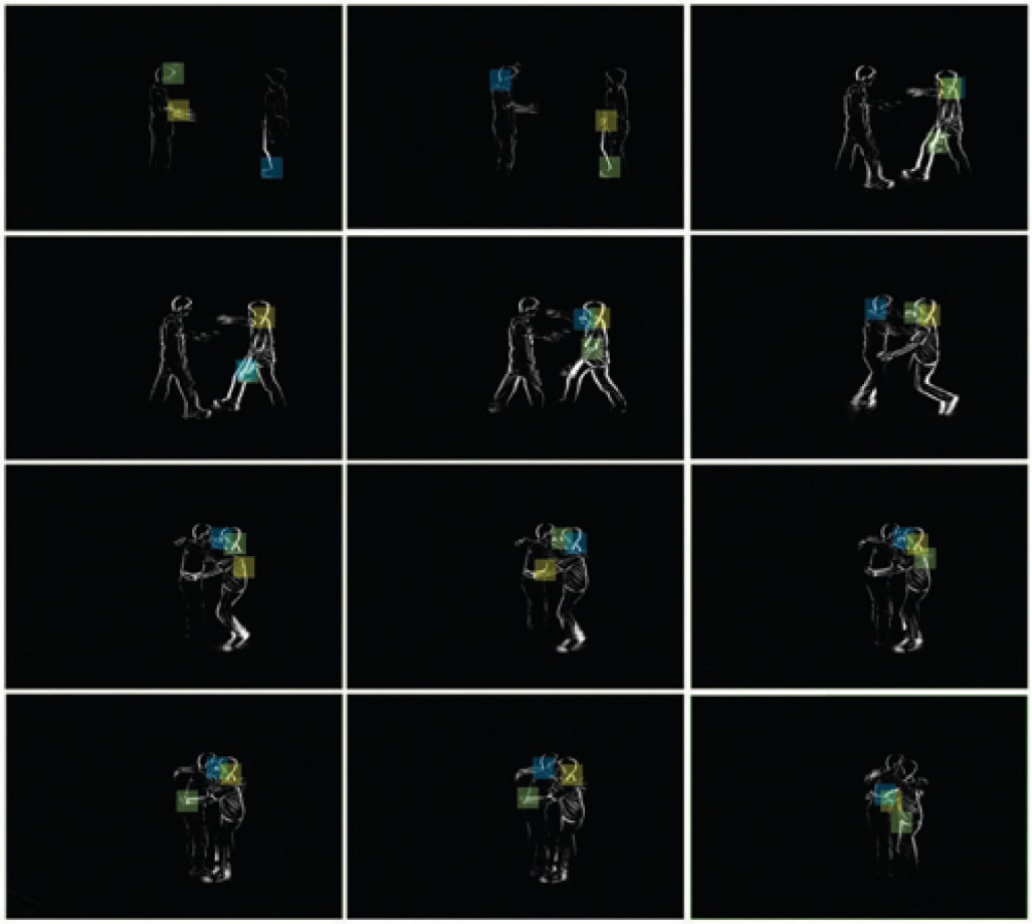


Figure 4. Illustration of the location of the extracted spatial-temporal interest points for action 'hug'. The highlighted points indicated the interest points of each frame.

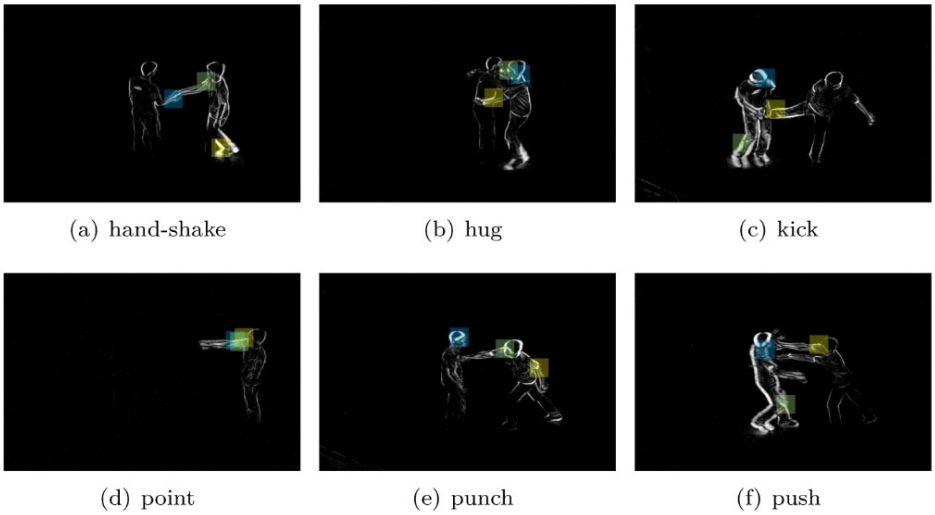


Figure 5. Cuboids generated around interest points for each action.

As seen in Figure 6. The cuboid (spatial temporal video patch) is extracted around each interest point and contains spatio-temporally windowed pixel values.

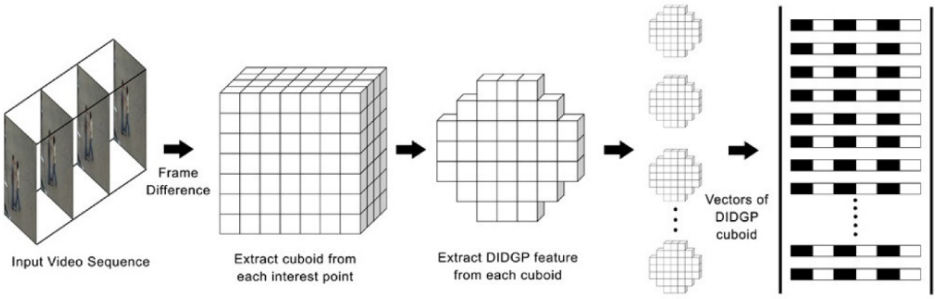


Figure 6. Extraction of raw features from a video sequence.

The size of cuboids in space-time volumes is set to four sizes (i.e., 49 (space) $\times$ 49 (space) $\times$ 14 (time), $49 \times 49 \times 21$, $49 \times 49 \times 28$, and $49 \times 49 \times 35$), in detection. By using this information of each cuboid, it is easy to describe and build a valid action recognition model.

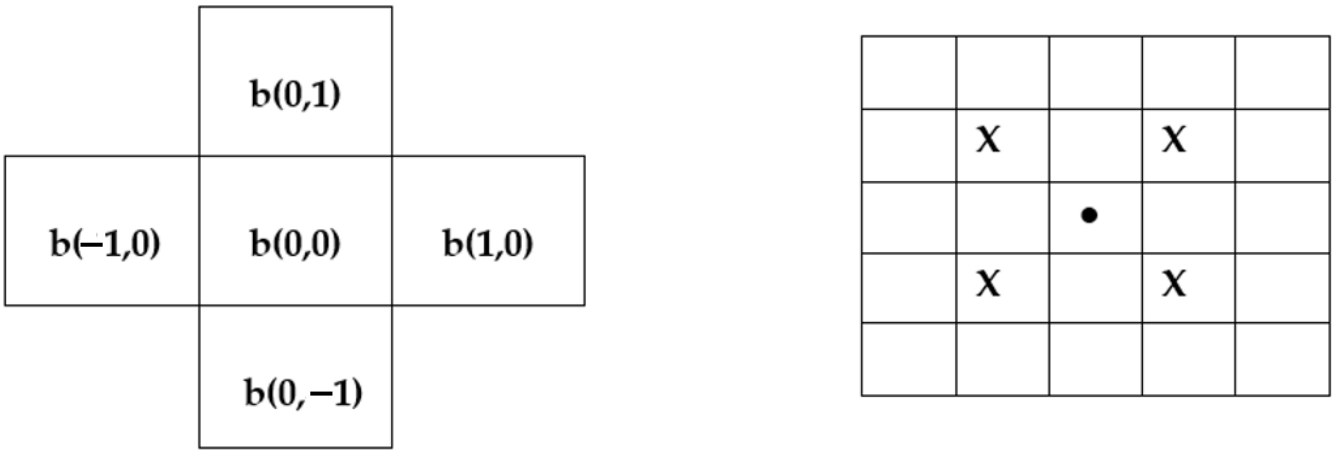
3.3. Feature Extraction Procedure

Once the cuboids of spatio-temporal interest points are identified, feature extraction is performed, and various features such as Difference Intensity Distance Group Patterns (DIDGP) in 2D/3D, Discrete Cosine Transform derivatives (DCT), and hybrid DWT+DCT are evaluated. As discussed above, the center block $b(0,\ 0)$ is kept at the interesting point detected using the procedure discussed in Section 3.3.1. A sampled patch size of 49×49 is positioned on the interest point and the DIDGP features are extracted.

3.3.1. Distance Relationship Calculation

This approach finds the distance relationship between the two blocks in the 7×7 block area of the extracted cuboid. Distance is calculated with the help of the center block $b(0,\ 0)$ is also called as the reference block. The concept of neighbouring pixels is applied to the blocks. This block $b(x,\ y)$ at $b(0,\ 0)$ consists of two horizontal and vertical neighbors, given as $(x+1, y)$, $(x-1,\ y)$, $(x,\ y+1)$, $x,\ y-1)$ in positions, $b(1,\ 0)$, $b(-1,\ 0)$, $b(0,\ 1)$, and, $b(0, -1)$ correspondingly as shown in Figure 7a,b. Every block is at b unit distance as seen in Figure 7a. The four diagonal neighbors of $b(x,\ y)$ at $b(0,\ 0)$ is given by,

$$(x+1,\ y+1),\ (x+1,\ y-1),\ (x-1,\ y-1),\ (x-1,\ y-1) \tag{4}$$



(**a**) (**b**)

Figure 7. (**a**) Unit distance representation of the block. (**b**) Diagonal neighbors at $\sqrt{2}$ from b.

They are at Euclidean distance of $\sqrt{2}$ from $b(0,\ 0)$.

In this work, the sampled patch size is $M = N = 49$ and each patch is divided into a 7×7 grid with $m = n = 7$ in pixels.

3.3.2. Distance Group Pattern

It is denoted by the different pattern groups organized by their distance from the central block as shown in Figure 8. The various distance group patterns obtained are denoted as $d = 1,\ d = \sqrt{2},\ d = 2,\ d = \sqrt{5},\ d = \sqrt{8},\ d = 3,\ d = \sqrt{10}$.

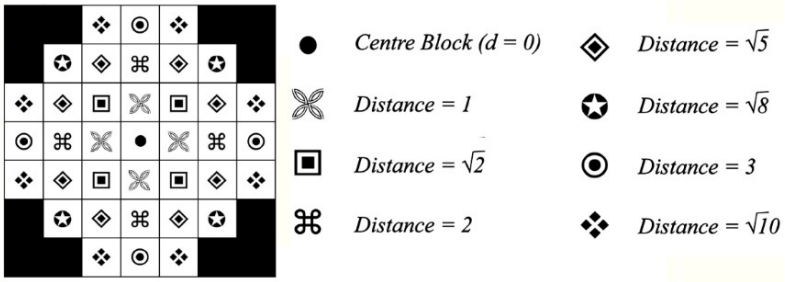


Figure 8. Distance Group Pattern.

To summarize, block distance is calculated from the central block. The eight different distance group is computed from 7×7 sub block region. The two dominant and different edge directions in a local region are defined as a corner. An interest point has a well-defined position in an image having local intensity maximum or minimum where the curvature is locally maximal. Harris 2D was utilized and the features extracted improve the performance of the activity recognition approach. Harris corner detector uses 2D Gaussian filter and 1D Gabor filters in spatial and temporal directions respectively. A response value is given at every position. Initially, various experiments are performed to fix the number of interest points for computation purpose. The number of interest points is varied as $n = 2,\ 3,\ 5$, and 7. Good performance is obtained with $n = 5$ and increasing the number of interest points increases the computational complexity. Thus, for further analysis, the number of interest points is fixed as 5. The highlighted spatial temporal interest points correspond to local maxima of response function as shown in Figure 3 for the local patches placed on the detected interest points in an action sequence. The 2D/3D DIDGP features are extracted from these patches and the information contains at each patch is utilized to describe and build the valid model for action recognition.

3.3.3. Signal Transformation Descriptors

The technique of time-frequency transformation helps in the conversion of the signal into various frequency components, making the features more accurate for the representation of the action. Further, it enriches the recognition ratio, and it has extensive usage of the image processing applications.

The discrete cosine transform (DCT) [35] characterizes an image data, sum of sinusoids of shifting magnitude, and frequency coefficients that are encoded individually for compression efficiency without dropping the information. It is mainly useful in the area of image processing, has many good properties like decorrelation and Energy compaction that removes the redundancy between neighboring pixels and discards the coefficients with relatively small amplitudes.

$$g(u,v) = \alpha(u)\alpha(v) \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cos \frac{(2x+1)u\pi}{2M} \cos \frac{(2y+1)v\pi}{2N} \tag{5}$$

$$\alpha(u) = \begin{cases} \sqrt{\frac{1}{M}}, & u = 0 \\ \sqrt{\frac{2}{M}}, & 1 \leq u \leq M-1 \end{cases} \tag{6}$$

$$(v) = \begin{cases} \sqrt{\frac{1}{N}}, & v = 0 \\ \sqrt{\frac{2}{N}}, & 1 \leq v \leq N-1 \end{cases} \tag{7}$$

where $u = 0,\ 1,\ 2, \ldots, M-1$ and $v = 0,\ 1,\ 2,\ \ldots,\ N-1$.

The 2D basis functions can be generated by multiplying the horizontally oriented 1-D basis functions with vertically oriented set of the same functions. The basic functions for $N = 8$. The basis functions exhibit a progressive increase in frequency both in the vertical and horizontal direction. The top left basis function assumes a constant value and is referred to as the DC coefficient. All other transform coefficients are called the AC Coefficients. The DCT is applied to the entire image to obtain the frequency coefficient matrix of the same dimension. In general, the DCT coefficients are divided into three bands namely low frequencies, middle frequencies and high frequencies. A transformation scheme must have the ability to pack input data into as few co-efficient as possible which allows the quantize to discard coefficients with relatively small amplitudes without introducing visual distortion in the reconstructed image. DCT exhibits excellent energy compaction for highly correlated images. The uncorrelated image has its energy spread out whereas the energy of the correlated image is packed into the low frequency region which is the top left region. The conventional coefficient selection approaches select the fixed elements of the DCT coefficients matrix. For an M $\times$ N size action frame. The feature extraction consists of two phases. In the first phase, the DCT is applied to the entire frame to obtain the DCT coefficients; a deterministic approach called Zigzag is used for coefficient selection of features in our work. Dimension of the DCT coefficient matrix is the same as the input frame.

The discrete wavelet transform (WT) has gained widespread acceptance in signal processing and image compression. The 'wavelets' are different from Fourier analysis that avoids the sine and cosine transform. This appears to be part of harmonic analysis of the wavelet family; it breaks the signal into a series of essential functions and models the complex phenomena effectively. The wavelet transition is independently determined at various frequencies for different segments in their respective time intervals. Multi-resolution analyses are designed in such a way that provides good time resolution and poor frequency resolution in high frequencies. In turn, proper frequency resolution and low time resolution result in high-frequency components for a short period and poor frequency components for a long duration. Here, wavelets are produced in the form of a single

prototype wavelet *y(t)* called mother wavelet which includes dilations and shifting and the group of wavelets resulting from (*x*) [39].

$$\Upsilon_{a,b}(x) = \frac{1}{\sqrt{a}}\Upsilon\left(\frac{x-b}{a}\right),\ (a, b\ \in R,\ a > 0) \tag{8}$$

where a is the scaling parameter and b is the shifting parameter. The mapping of discrete wavelet transform is given as

$$(Tf)_{a,b} = \int f(x)\ \Upsilon_{a,c}(x)dx \tag{9}$$

where the mother wavelet $\Upsilon_{a,b}(x)$ satisfies the below equation:

$$\int \Upsilon(x)dx = 0 \tag{10}$$

DWT is used to transforming the gray-scale images to spatial and frequency domain at the same time, where x is any signal which wavelet decomposition to be performed, q_l and q_h are low pass and high pass filters with half the cut-off frequency from the previous one. Such a transformation is applied recursively on the low-pass series until the desired number of iterations is reached. In frequency domain, when the facial image is decomposed using two-dimensional wavelet transform, four sub regions are obtained. These regions are: one low-frequency region LL (approximate component), and three high-frequency regions, namely, LH (horizontal component), HL (vertical component), and HH (diagonal component), respectively. The LL image is generated by two continuous low-pass filters; HL is filtered by a high- pass filter first and a lowpass filter later; LH is created using a low-pass filter followed by a high-pass filter; and HH is generated by two successive high-pass filters. Subsequent levels of decomposition follow the same procedure by decomposing the LL sub image of the previous level. Since the LL part contains most important information and discards the effect of noises and irrelevant parts, the LL part is adopted for further analysis. In the proposed work, two-level 2D discrete wavelet decomposition is performed on the motion images. The Dimension of the DWT coefficient matrix is the same as the input frame.

Hybrid DWT+DCT. The process for feature extraction using hybrid DWT+DCT is as follows: input the motion frames of size M × N is applied with 2D-DWT to obtain LL band; further, the LL band is divided into 32 narrow-width bands. Each band is of size 16. It is observed that dividing the image into 32 bands gives better results in terms of recognition accuracy. We apply the 2D-DCT on each band to obtain the fewer number of DCT coefficients with high compaction energy on lower frequency which gives dominant magnitudes are obtained by arranging them in descending order in each band of the image. The first dominant magnitudes in each band contain different characteristics in comparison to other magnitudes in the respective bands. All hybrid DWT-DCT coefficients pertaining to first dominant magnitudes in each band of the motion frame are considered as the feature vector.

3.4. Principal Component Analysis (PCA)

PCA is a powerful and significant statistical approach in broad areas, and it is used in high-ceiling dimension to identify patterns. PCA 'combines' the attribute spirit by producing a replacement for a smaller collection of variables [40,41]. The original data are projected as a reduced set. Assume that $x_1,\ x_2,\ \ldots,\ x_p$ are P training vectors, each fit into one of the N classes $\{\zeta_1,\ \zeta_2,\ \ldots,\ \zeta_N\}$. Hence, the training vector, x_p, can also be projected in a lower dimension vector space y_p, using a linear transform of orthonormal form and given as $y_p = W^T x_p$. The eigenvalues and eigenvectors help in the construction of the

transformation matrix (W) and it includes the covariance matrix ($\sum$) of the input data. By default, input data covariance can be calculated as

$$\sum = \frac{1}{P}\sum_{p=1}^{P}(x_p - \mu)(x_p - \mu)^T, \tag{11}$$

Here, μ is considered as the mean vector of all the sample training images.

The covariance matrix has the eigenvectors $e_1, e_2, \ldots, e_K$ associated with the eigen values $\lambda_1 \geq \lambda_2 \geq \ldots, \geq \lambda_K$, correspondingly; here K, is defined as the feature vector dimension. The eigenvectors relating to D concentrated eigenvalues, i.e., $(W) = [e_1 e_2 \ldots e_D]$, are accomplished by holding D(D $\ll$ K) in the transformation matrix (W). Here, the direction (e_i) of relevant feature and the information of population arises with the increased range of spread (variance) and feature $[e_1 e_2 \ldots e_D]$ retain the substantial portion of the feature population data for any D.

The projection of lower dimension space (say vector t) for a given specific sample of test data is decreased using distance match algorithm, and the corresponding class is assigned to the feature training vector X_{i_o}, where $i_o = \arg\min_{1 \leq i \leq p} || t - y_i ||$, where $||.||$ indicates the Euclidean distance in R D. The initial eigenvector indicates the two-dimensional data and represents the direction of maximum zero variance. The second eigenvector is orthogonal to the first, and it relates to the next maximum direction of variance.

In this work, we utilized the Principal Component Analysis (PCA) approach for linear feature extraction used for unsupervised feature selection based on eigenvectors analysis to identify critical original features for principal component. This can dramatically impact the performance of machine learning.

3.5. Classification Methods

In this work, the SVM classifier and Random Forest classifier are used to test the efficacy of the classifier in the UT-Interaction dataset. The classifiers applied are as follows:

3.5.1. Support Vector Machines

Support Vector Machine (SVM) is a widely used approach for classifying visuals in pattern recognition [42,43]. It achieves greater success in the performance of optimization theory by using vital pattern recognition [44], and it mainly involves in the practice of kernel learning algorithm. Typically, the classification task involves sample training and testing data. The separation of training data is given by $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ into two classes, where $x_1 \in R_N$ resembles the feature vector of n-dimensional and $y_i = \{+1, -1\}$ is the required class labels. SVM aims to forecast the target values from the testing set with the help of generating a model. The hyper plane in binary classification is $w \cdot x + b = 0$, where $w \in R^n$, $b \in R$ is involved in the separation of two classes in the distinct space Z [45]. $M = \frac{2}{||w||}$ is assumed as the maximum margin as shown in Figure 9. Lagrange multipliers help in resolving the minimization problem $\alpha_i (i = 1, 2, \ldots, m)$ inturn w and b are considered as the most optimum values gained from Equation (12).

$$f(x) = sgn\left(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b\right) \tag{12}$$

In order to optimize the margin and minimize the training error, the non-negative slack variables ξ_i are used. The soft margin classifier is acquired by optimizing the Equations (13) and (14).

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C\sum_{i=1}^{l} \xi_i \tag{13}$$

$$y_i(w^T \phi(x_i) + b \geq 1 - \xi_i, \ \xi_i \geq 0 \tag{14}$$

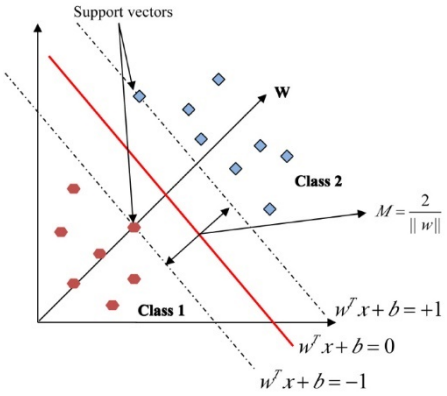


Figure 9. Illustration of hyperplane in linear SVM.

If, in case the sample training, data are not linearly separable, then the input space with the kernel function is mapped to the high dimensional space $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ [42]. Some of the main characteristics of kernel functions are listed in Table 1. From Table 1, γ and r are known as parameters of inner kernel.

Table 1. SVM inner product kernel types.

Types of Kernel Inner Product Kernel
Linear $K(x_i, x_j) = x^T{}_i x_j$
Polynomial $K(x_i, x_j) = (\gamma x^T{}_i x_j + \gamma)^d$, $\gamma > 0$
Radial Basis Function $K(x_i, x_j) = \exp(-\gamma x_i - x_j{}^2)$, $\gamma > 0$
Sigmoid $K(x_i, x_j) = \tanh(\gamma x^T{}_i - x_j + r)$

Multiclass SVM helps in the construction of N-binary classifier, and one class is isolated from the rest. The ith class training sets comprises of positive labels, and all other labels are negative. The ith SVM unravels the ith decision function given in Equation (12).

A grid search technique is used in the RBF kernel to predict the value of the C and γ parameters in LIBSVM space [45]. The ideal value parameters, such as C and γ, are not previously known, and it helps in the optimal performance of the classifier, where C refers to the variable slack and γ defines the decision boundary curvature or penalty parameter.

3.5.2. Random Forest

Leo Breiman's [46] enhanced Random Forest has a collection of non-pruned classification trees made up of a random selection of trials of training data samples. Here, the induction process helps in the selection of random features. Aggregation attains the majority vote by predicting in case of classification, and it undergoes averaging for regression. Breiman's Random Forest approach combines random subspace elements and bagging uniquely to use decision trees as the base classifier. This approach is not suitable for handling a large number of dissimilar features, and, hence, overfitting occurs in noisy classification/regression tasks in distinct data sets. In this work, the maximum number of trees is 100 and the depth is fixed to 50.

4. Experimental Setup

The evaluation of the proposed work is assessed using the UT-Interaction dataset on Set 1 and Set 2 (Figure 10). The experiments are performed with MATLAB R2019 in Windows 10 operating system on a computer with Intel i7 Processor having 16 GB RAM. The extraction of cuboids in the feature detection step is carried by interest points, that access the necessary information needed for transform-based features and DIDGP features. Moreover, this model is tested with real time benchmark-video sequences and 10 frames per second are required for processing. When we increase the computation power to GPU based systems, this algorithm can work in speed of 25 to 30 frames per second.

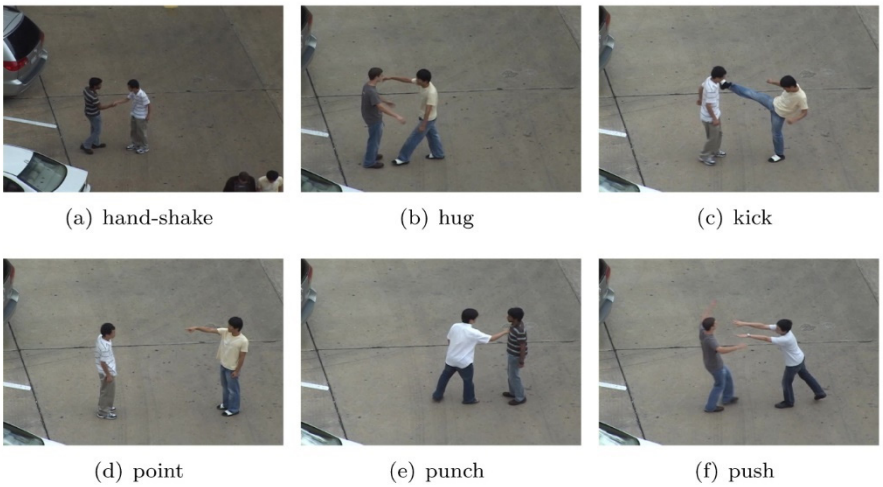


Figure 10. Example frames of each action of UT-Interaction dataset.

4.1. 3D-DIDGP

The spatio-temporal information generated within each cuboid and the concatenated vector is identical to the sub-block number. Hence, the descriptor vectors are assigned to a lower-dimensional space in the cuboid times.

4.2. 2D-DIDGP

The spatio-temporal information generated within each interest point DIDGP descriptor is concatenated and expected to project in lower-dimensional space through descriptor vectors.

4.3. DCT

In the DCT feature detection step, the 2D windowed pixel that is extracted within the spatio-temporal interest point of the difference image has the length of $49 \times 49 = 2401$. From this, it is divided into $8 \times 8 = 64$ sub blocks were achieved, then DCT method is applied on 8×8 sub blocks to obtain the complex coefficients. The extracted descriptors of length 64 are given to lower dimensional space by varying the step size of 5 to achieve recognition accuracy.

4.4. DWT

Daubechies wavelet transform can be applied to separates finite data into different frequency components, such as approximation coefficients matrix cA and other coefficients matrices cH (horizontal), cV (vertical), and cD (diagonal). Similar to the Discrete Cosine Transform, DWT is used, and results in the smallest number of coefficients of descriptor length 48.

4.5. Hybrid DWT+DCT

In this method, the process of DWT and DCT act together. Where the input is DCT coefficients, it helps in minimizing the redundancy to discriminate against the action more efficiently. Finally, the classification process involves leaving the one-out-cross validation (LOOCV) method for performance assessment of the non-linear support vector machine (SVM) with the RBF kernel. Here the best parameters are chosen by 10-fold cross-validation in a grid search on the sample training data, and Random Forest also uses the LOOCV approach for performance evaluation.

4.6. Dataset

The UT-Interaction dataset contains a video sequence of six classes of human–human interaction classes: shake-hands, dot, hug, push, kick and punch. There are a total of 20 video sequences of approximately 60 s in length. Each video includes two experiences,

providing an average of eight human activities per video. More than 15 different styles of participants with various clothing requirements perform in the videos. The images are taken at a resolution of 720 × 480 and 30 fps, and one person's height is approximately 200 pixels in the video. The video is divided into two parts. Set 1 consists of 10 video sequences pictured in a parking lot with modifying zoom rates and their backgrounds are almost static with the minimal camera jitter. Set 2 (i.e., 10 additional sequences) is taken on a windy day lawn with a slightly shifting background with additional camera jitters. Every package has a different context, size, and lighting.

4.7. Evaluation Metrics

Table 2 illustrates a confusion matrix for a human emotion recognition problem having true positive (*TP*), false positive (*FP*), true negative (*TN*), and false negative (*FN*) class values [47,48]. If the classifier predicts a correct response of class at each instance, it is counted as "success"; if not, it is an "error". The overall performance of the classifier is obtained by error rate, which is a proportion of the errors made over the whole set of instances.

Table 2. Confusion matrix for classification.

Actual Values / **Predicted Values**	**Positive**	**Negative**
Positive	*TP*	*FP*
Negative	*FN*	*TN*

From the confusion matrix it is possible to extract a statistical metrics (Precision, Recall, and F-measure) for measuring the performance of classification systems and is defined as follows: Precision (*P*) or detection rate is a ratio between correctly labelled instances and total labelled instances. It is a percentage of positive predictions in specific class that are correct and it is defined by:

$$precision\ (P) = \frac{TP}{TP + FP} \tag{15}$$

where, *TP* and *FP* are the number of true positive and false positive predictions for the particular class.

Recall (*R*) or Sensitivity is a ratio between correctly labeled instances and total instances in the class. It has an ability to measure the prediction model and is also called as true positive rate. It is defined by:

$$Recall\ (R) = \frac{TP}{TP + FN} \tag{16}$$

where, *TP* and *FN* are the numbers of true positive and false negative predictions for the particular class. *TP* + *FN* is the total number of test examples of the particular class.

The *F*-measure is the harmonic mean of precision and recall and it attempts to give a single measure of performance. A good classifier can provide both recall and precision values high. The F-measure is defined as:

$$F_\beta = \frac{(1+\beta)^2 \cdot TP}{(1+\beta)^2 \cdot TP + (\beta)^2 \cdot FN + FN} \; or \; F_1 = 2 \cdot \frac{F \cdot R}{P + R} \tag{17}$$

where β is the weighting factor. Here, $\beta = 1$, that is, precision and recall are equally weighted and used to measure the F_β-score which is also known as F_1—measure.

5. Experimental Results

5.1. Transform Based Descriptor

A modern discriminative feature descriptor approach of the time-frequency transformation technique is used to identify the features in a more reliable way using action representation. It also increases the detection rate which is commonly used in the field of image processing. Here various transformation methods like Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT), as well as the combination of DCT and DWT (Hybrid DWT+DCT), which helps in utilizing human action recognition due to its excellent performance in image and video processing.

Figure 11 illustrates the performance comparison of transform-based descriptor methods, using the RBF-kernel SVM and Random Forest, which shows the results classified by SVM from Figure 11a,b. The results for Set 1 and Set 2 are shown in Figure 11c,d by using the Random Forest Classifier. It appears that the Hybrid DWT+DCT approach outperforms both DCT and DWT techniques, and the Hybrid DWT+DCT results on both sides are 81.04% and 71.69%, with the size 35, respectively. The accuracy of the identification increases as the dimension increases to a certain amount. Furthermore, it was found that further dimensional increases do not improve. The SVM classification gives greater accuracy to recognize than the Random Forest and for the SVM classified DWT+DCT method, the results of the Random Forest results for both sets of UT-Interaction are approximately 17% to 20%.

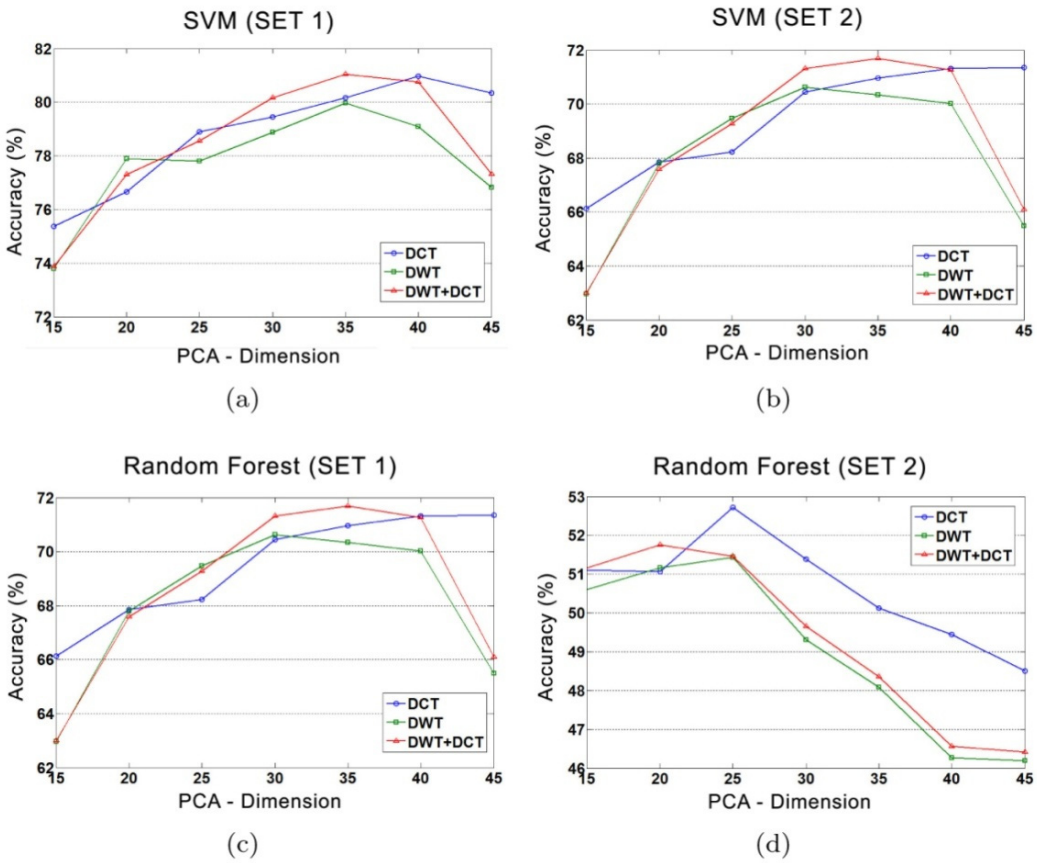


Figure 11. Comparison of accuracy under different sets (**a**,**b**) show accuracy of SVM with RBF kernel against PCA with different dimensions; (**c**,**d**) give accuracy of Random Forest against PCA with different dimensions.

5.2. 2D-DIDGP

The 2D-DIDGP features are extracted on each spatial-temporal interest point. PCA undergoes dimensionality reduction and the resultant features are supplied to the SVM and the Random Forest Classifier to achieve a good description. Figure 12 displays the

results of the projected 2D-DIDGP. Figure 12a exhibits the SVM results with RBF kernel on the UT-Interaction dataset (Set 1 and Set 2). From this average performance accuracy, 89.87% and 87.59% were obtained on 2D-DIDGP at the dimension of 18 on Set 1 and Set 2, accordingly. The performance improves up to a certain point and then decreases slightly. From this, the best results were achieved with a dimension of 18. Figure 12b shows the results of the Random Forest classifier, the obtained performance of 87.83% and 86.18% on 2D-DIDGP at the dimension of 15 on Set 1 and Set 2 respectively.

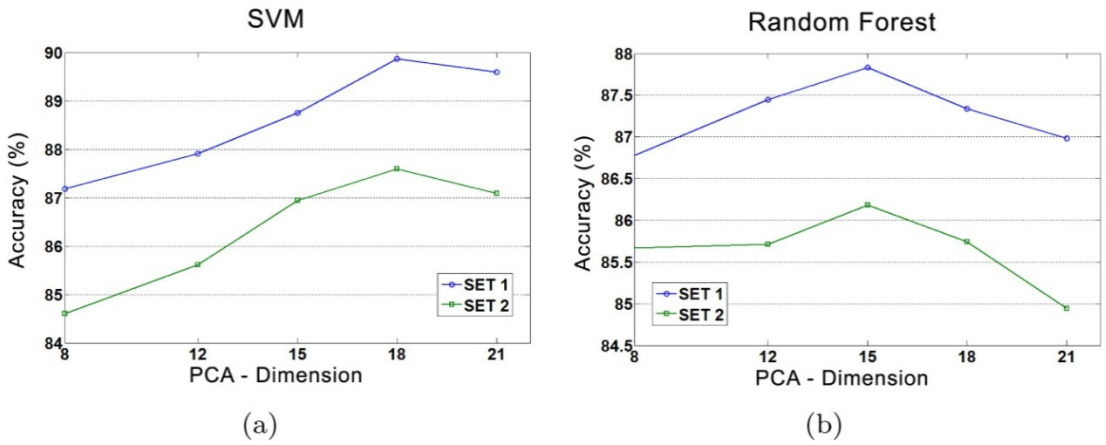


Figure 12. (**a**,**b**) Show accuracy of SVM with RBF kernel and Random Forest against PCA with different dimensions on UT-Interaction dataset (Set 1 and Set 2).

Tables 3 and 4 show the classification results of the UT-Interaction dataset with RBF kernel using SVM on Set 1 and Set 2 and the correct response lies along with the diagonal entries of the table. Most classes are predicted well, with some confusion between punch and push on Set 1. In Set 2, the results are vaguely predicted well and show that the action point and punch are likely to be confused with push. From both cases, the hand-based action (point, punch, and push) is the same and, therefore, the confusion is expected due to the similarity of the poses.

Table 3. Confusion matrix (%) of 2D-DIDGP on Set 1 using SVM with RBF kernel function, where A = hand-shake, B = hug, C = kick, D = point, E = punch, F = push.

	A	**B**	**C**	**D**	**E**	**F**
A	**92.66**	1.68	1.68	1.22	0.61	2.14
B	1.21	**91.68**	2.05	1.57	0.72	2.77
C	2.23	2.75	**88.73**	1.83	1.83	2.62
D	2.48	3.73	1.45	**90.06**	1.04	1.24
E	2.61	2.43	2.43	0.93	**84.89**	6.72
F	1.57	3.13	1.72	0.63	1.72	**91.22**

Table 4. Confusion matrix (%) of 2D-DIDGP on Set 2 using SVM with RBF kernel function, where A = handshake, B = hug, C = kick, D = point, E = punch, F = push.

	A	**B**	**C**	**D**	**E**	**F**
A	**90.06**	5.25	0.55	0.14	1.52	2.49
B	2.78	**89.52**	1.26	0.76	1.89	3.79
C	3.37	4.16	**85.94**	0.59	1.39	4.55
D	1.99	1.99	2.33	**83.72**	2.66	7.31
E	2.73	4.92	0.36	0.36	**85.79**	5.83
F	1.69	3.38	1.52	0.34	2.54	**90.52**

Tables 5 and 6 show the UT-Interaction dataset confusion table on Set 1 and Set 2 using the Random Forest; one can observe that the action punch is confused with hug and push. In Set 2, most actions are not classified well, except handshake.

Table 5. Confusion matrix (%) of 2D-DIDGP on Set 1 using Random Forest classifier.

	A	B	C	D	E	F
A	**90.83**	4.59	1.99	0.61	0.92	1.07
B	2.90	**90.83**	2.65	0.60	0.97	2.05
C	2.23	3.93	**89.91**	0.92	0.79	2.23
D	5.18	4.97	2.69	**85.09**	0.62	1.45
E	2.61	8.02	1.87	1.68	**81.90**	3.92
F	2.04	6.11	2.66	0.31	0.47	**88.40**

Table 6. Confusion matrix (%) of 2D-DIDGP on Set 2 using Random Forest classifier.

	A	B	C	D	E	F
A	**90.19**	6.35	1.10	0.00	1.24	1.10
B	6.06	**88.51**	1.01	0.13	1.64	2.65
C	3.17	6.93	**80.00**	0.79	3.76	5.35
D	2.99	3.32	1.99	**84.72**	1.99	4.98
E	1.09	6.74	0.73	0.18	**86.70**	4.55
F	1.52	6.09	2.03	0.51	2.88	**86.97**

5.3. 3D-DIDGP

The obtained 3D-DIDGP cuboid features were extracted from each spatiotemporal interest point. Reduction of dimensionality with PCA by varying step size of 15 and projected features are fed to the SVM and Random Forest classifiers to achieve fair comparison. Figure 13 displays the results of projected 3D-DIDGP under different cuboid sizes in times (14, 21, 28, and 35). Figure 13a,b show the results of SVM with RBF kernel on UT-Interaction dataset (Set 1 and Set 2). From this average performance accuracy, 96.32% and 92.03% were obtained on 3D-DIDGP28 at the dimension of 30 on Set 1 and Set 2 respectively.

The accuracy hardly improves as the number of cuboid size increase. Figure 13c,d show the results of the Random Forest classifier, the obtained better performance of 90.86% and 82.60% on 3D-DIDGP28 at the dimension of 30 on Set 1 and Set 2, respectively. The cuboid size of 28 outperforms the other cuboid sizes and the best result is achieved on a projected dimension of 30 on both classifiers.

The recognition capability of the 3D-DIDGP method on SVM with RBF kernel demonstrated by confusion matrix for Set 1 and Set 2 are given in Tables 7 and 8, respectively. In Set 1, 3D-DIDGP has outstanding recognition competence in point, punch, push, and hug. However, for a handshake, hug, and kick, results are somewhat confusing. In Set 2, confusion matrix results of point and push have good recognition capability and show that the action kick, hug, and handshake are likely to be confused with push. This is because in the UT-interaction dataset, a certain sequence like push and punch are correlated, and it is very complex to differentiate with human eyes.

Table 7. Confusion matrix (%) of 3D-DIDGP28 on Set 1 using SVM classifier with RBF kernel function.

	A	B	C	D	E	F
A	**94.66**	0.76	1.27	1.02	2.29	0.00
B	0.21	**95.47**	1.03	0.41	2.88	0.00
C	2.22	0.49	**92.84**	0.74	3.46	0.25
D	0.00	0.00	0.00	**100.00**	0.00	0.00
E	0.65	0.86	1.29	0.22	**96.77**	0.22
F	0.46	0.23	0.92	0.00	0.23	**98.16**

Table 8. Confusion matrix (%) of 3D-DIDGP28 on Set 2 using SVM classifier with RBF kernel function.

	A	B	C	D	E	F
A	**91.51**	0.63	0.94	0.63	1.57	4.72
B	1.33	**88.67**	1.67	0.67	3.33	4.33
C	1.37	2.73	**82.24**	0.55	2.46	10.66
D	0.00	0.00	0.00	**100.00**	0.00	0.00
E	0.68	0.46	0.91	1.14	**93.15**	3.65
F	1.57	0.79	0.26	0.00	0.79	**96.59**

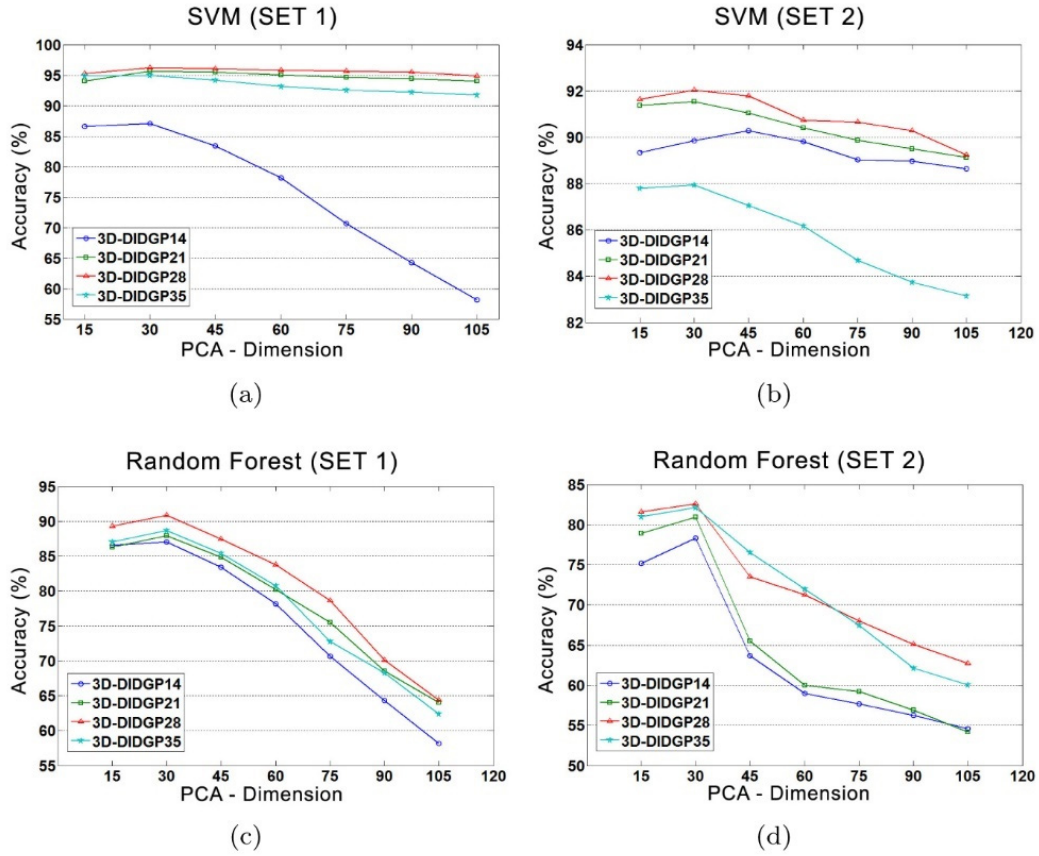


Figure 13. Comparison of accuracy under different conditions (**a**,**b**) show accuracy of different Cuboid size on SVM against PCA with different dimensions. (**c**,**d**) give accuracy of different Cuboid size on Random Forest against PCA with different dimensions.

Tables 9 and 10 display the results classified by Random Forest. The appropriate responses to the main diagonal appear in Table 9, where the majority of actions were appropriately categorized but the discrimination on a handshake and kick actions on Set 1 was comparatively less effective. Table 10 shows the results on Set 2, most of the actions are not predicted well except point action, the greatest confusion between hug, kick, handshake, push and punch were found difficult to segregate reliably. The performance of the 3D-DIDGP method on SVM with RBF kernel is outperformed by 5.46% and 9.43% on Set 1 and Set 2, respectively, by the Random forest classifier.

Table 9. Confusion matrix (%) of 3D-DIDGP28 on Set 1 using Random Forest classifier.

	A	B	C	D	E	F
A	**89.57**	2.54	5.34	0.51	1.27	0.76
B	1.03	**93.21**	3.50	0.41	1.23	0.62
C	7.16	7.16	**79.01**	0.99	2.72	2.96
D	0.00	0.62	0.00	**99.38**	0.00	0.00
E	0.86	1.72	2.15	0.86	**91.83**	2.58
F	1.84	1.84	2.53	0.00	1.61	**92.18**

Table 10. Confusion matrix (%) of 3D-DIDGP28 on Set 2 using Random Forest classifier.

	A	B	C	D	E	F
A	**77.67**	2.20	1.26	0.63	8.81	9.43
B	6.33	**67.67**	2.33	0.33	14.67	8.67
C	3.28	3.28	**75.96**	0.82	13.11	3.55
D	0.00	0.00	0.00	**100.00**	0.00	0.00
E	1.14	1.37	3.65	0.68	**89.50**	3.65
F	2.62	2.36	2.36	0.00	7.87	**84.78**

6. Performance Analysis of Different Methods

The overall results reported by SVM and the Random Forest classifiers are shown in Figure 14a,b. The 3D-DIDGP approach succeeded the methods based on 2D-DIDGP and Transform, with comparable performance. For the 3D-DIDGP process, the result reported by SVM exceeds the results of Random Forest by about 7–15%. The findings are not limited to random forests.

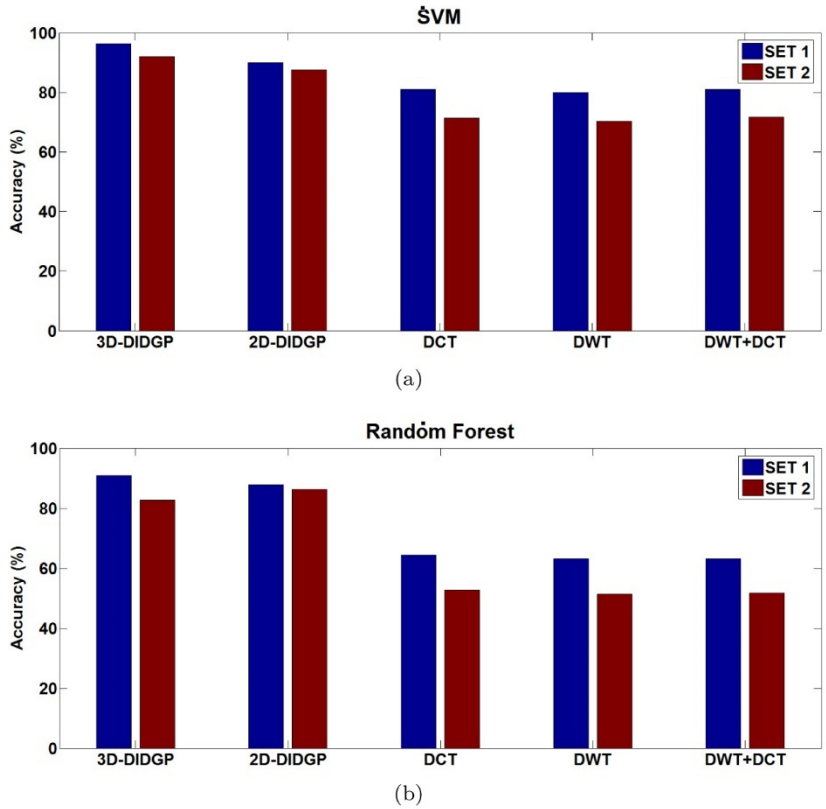


Figure 14. Comparison of classification accuracies of proposed methods on UT-Interaction dataset (Set 1 and Set2) (**a**,**b**) shows the accuracy of SVM and Random Forest with PCA on UT-Interaction dataset Set 1 and Set 2, respectively.

To measure the efficiency of the proposed method, the results obtained by the proposed method are quantitatively compared with the most advanced results, and compared them are shown in Table 11. Based on the comparison, the proposed method shows good results on the UT-Interaction dataset. The experimental results validate their accuracy and efficiency in human action recognition and also indicate the potential of proposed techniques. Moreover, the results indicate that the 3D-DIDGP is quite promising and convincing for human action recognition in surveillance videos as it extracts reliable feature information when compared to 2D-DIDGP and transforms-based descriptors. The 3D-DIDGP based method outperforms the state-of-the-art recognition algorithms and obtained 96.32% and 82.03% for UT-Interaction dataset (Set 1 and Set 2).

Table 11. State-of-the-art Recognition Accuracy (%) for the UT-Interaction (Set 1 and Set 2) datasets.

Method	Year	Set 1	Set 2
Proposed	-	96.32	92.03
Poonkodi et at. [49]	2020	96.08	89.3
Xiao-Fei et al. [50]	2017	94	-
Uddin et al. [51]	2017	91.67	-
Liang et al. [52]	2016	92.3	-
Zhang et al. [53]	2012	95	90
Xiaojiang et al. [32]	2013	94.5	91.7
Vahdat et al. [54]	2011	93	90
Motiian et al. [55]	2013	91.8	87.87
Waltisberg et al. [31]	2010	88	77
Ryoo et al. [56]	2010	85	70
Angela et al. [57]	2012	77	73

7. Conclusions

In this work, spatio-temporal interest points (STIPs) and transform-based descriptor methods are proposed for the application of the human action recognition task. Actions are described as sequences of successive human body poses. Initially, the approach captures the local motion of the human body poses from a video that is invariant to size and shape changes. The Difference Intensity Distance Group Pattern (DIDGP) approach yields an efficient and effective human action recognition system. Experiments are conducted on the UT-Interaction dataset considering different activities viz (handshake, hug, kick, point, punch, and push). The experimental results validate their accuracy and efficiency in human action recognition, and also indicate the potential of proposed techniques. Moreover, the results indicate that the 3D-DIDGP is quite promising and convincing for human action recognition in surveillance videos as it extracts reliable feature information when compared to 2D-DIDGP and transforms-based descriptors. The 3D-DIDGP based method outperforms the state-of-the-art recognition algorithms and obtained 96.32% and 82.03% for UT-Interaction dataset (Set 1 and Set 2). Future work will focus on efforts to extract the robust features to effectively understand human actions.

Author Contributions: Conceptualization, J.A., A.A.K. and S.T.; methodology, M.A.H., J.A. and R.D.; software, M.A., J.A. and L.V.; validation, M.I.A., R.K., L.V. and R.D.; formal analysis, I.K. and R.K.; investigation, J.A. and A.A.K.; resources, S.T. and M.A.H.; data curation, M.A.; writing—original draft preparation, I.K. All authors have read and agreed to the published version of the manuscript.

Funding: Mohamed Alwanain like to thank Deanship of Scientific Research at Majmaah University for supporting this work under Project No. R-2022-224. This research was supported by Researchers Supporting Program (TUMA-Project-2021–14), AlMaarefa University, Riyadh, Saudi Arabia. Moreover, this research was also supported by College of Computing, Khon Kaen University, Thailand.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: Mohamed Alwanain like to acknowledge Deanship of Scientific Research at Majmaah University for supporting this work under Project No. R-2022-224. The authors deeply acknowledge the Researchers Supporting Program (TUMA-Project-2021–14), AlMaarefa University, Riyadh, Saudi Arabia for supporting steps of this work. Also, we would like to acknowledge College of Computing, Khon Kaen University, Thailand.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Russo, P.; Ticca, S.; Alati, E.; Pirri, F. Learning to See Through a Few Pixels: Multi Streams Network for Extreme Low-Resolution Action Recognition. *IEEE Access* **2021**, *9*, 12019–12026. [CrossRef]
2. Xia, H.; Gao, X. Multi-Scale Mixed Dense Graph Convolution Network for Skeleton-Based Action Recognition. *IEEE Access* **2021**, *9*, 36475–36484. [CrossRef]
3. Park, J.; Cho, S.; Kim, D.; Bailo, O.; Park, H.; Hong, S.; Park, J. A Body Part Embedding Model with Datasets for Measuring 2D Human Motion Similarity. *IEEE Access* **2021**, *9*, 36547–36558. [CrossRef]
4. Gao, Z.; Guo, L.; Guan, W.; Liu, A.-A.; Ren, T.; Chen, S. A Pairwise Attentive Adversarial Spatiotemporal Network for Cross-Domain Few-Shot Action Recognition-R2. *IEEE Trans. Image Process.* **2020**, *30*, 767–782. [CrossRef]
5. Arunnehru, J.; KalaiselviGeetha, M. Automatic human emotion recognition in surveillance video. In *Intelligent Techniques in Signal Processing for Multimedia Security*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 321–342.
6. Thalapathiraj, S.; Baskaran, B.; Arunnehru, J. Novel approach for texture feature extraction and classification of satellite images using modified Hilbert matrix. *AIP Conf. Proc.* **2019**, *2112*, 020154. [CrossRef]
7. Moeslund, T.B.; Hilton, A.; Krüger, V. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.* **2006**, *104*, 90–126. [CrossRef]
8. Turaga, P.; Chellappa, R.; Subrahmanian, V.S.; Udrea, O. Machine Recognition of Human Activities: A Survey. *IEEE Trans. Circuits Syst. Video Technol.* **2008**, *18*, 1473–1488. [CrossRef]
9. Poppe, R. Vision-based human motion analysis: An overview. *Comput. Vis. Image Underst.* **2007**, *108*, 4–18. [CrossRef]
10. Poppe, R. A survey on vision-based human action recognition. *Image Vis. Comput.* **2010**, *28*, 976–990. [CrossRef]
11. Weinland, D.; Ronfard, R.; Boyer, E. A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.* **2011**, *115*, 224–241. [CrossRef]
12. Hur, T.; Bang, J.; Huynh-The, T.; Lee, J.; Kim, J.-I.; Lee, S. Iss2Image: A Novel Signal-Encoding Technique for CNN-Based Human Activity Recognition. *Sensors* **2018**, *18*, 3910. [CrossRef]
13. Cruciani, F.; Vafeiadis, A.; Nugent, C.; Cleland, I.; McCullagh, P.; Votis, K.; Giakoumis, D.; Tzovaras, D.; Chen, L.; Hamzaoui, R. Feature learning for Human Activity Recognition using Convolutional Neural Networks. *CCF Trans. Pervasive Comput. Interact.* **2020**, *2*, 18–32. [CrossRef]
14. Arunnehru, J.; Chamundeeswari, G.; Bharathi, S.P. Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos. *Procedia Comput. Sci.* **2018**, *133*, 471–477. [CrossRef]
15. Vaijayanthi, S.; Arunnehru, J. Synthesis approach for emotion recognition from cepstral and pitch coefficients using machine learning. In *International Conference on Communication, Computing and Electronics Systems*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 515–528.
16. Laptev, I. On space-time interest points. *Int. J. Comput. Vis.* **2005**, *64*, 107–123. [CrossRef]
17. Dollar, P.; Rabaud, V.; Cottrell, G.; Belongie, S. Behavior recognition via sparse spatio-temporal features. In Proceedings of the 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, 15–16 October 2005; pp. 65–72.
18. Shen, Y.; Foroosh, H. View-invariant action recognition using fundamental ratios. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–6.
19. Natarajan, P.; Singh, V.K.; Nevatia, R. Learning 3D action models from a few 2D videos for view invariant action recognition. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2006–2013.
20. Wang, Y.; Sabzmeydani, P.; Mori, G. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In *Human Motion–Understanding, Modeling, Capture and Animation*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 240–254.
21. Fathi, A.; Mori, G. Action recognition by learning mid-level motion features. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
22. Elgammal, A.; Shet, V.; Yacoob, Y.; Davis, L. Learning dynamics for exemplar-based gesture recognition. In Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Proceedings, Madison, WI, USA, 18–20 June 2003; Volume 1, pp. 571–578.

23. Thurau, C.; Hlavac, V. Pose primitive based human action recognition in videos or still images. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
24. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local svm approach. In Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, 26 August 2004; Volume 3, pp. 32–36.
25. Shechtman, E.; Irani, M. Space-time behavior based correlation. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 405–412.
26. Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning realistic human actions from movies. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
27. Gorelick, L.; Blank, M.; Shechtman, E.; Irani, M.; Basri, R. Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 2247–2253. [CrossRef]
28. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. Hmdb: A large video database for human motion recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.
29. Marszalek, M.; Laptev, I.; Schmid, C. Actions in context. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 2929–2936.
30. Ryoo, M.S.; Aggarwal, J.K. UT-Interaction Dataset, ICPR Contest on Semantic Description of Human Activities (SDHA). 2010. Available online: https://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html (accessed on 17 July 2022).
31. Waltisberg, D.; Yao, A.; Gall, J.; van Gool, L. Variations of a hough-voting action recognition system. In *Recognizing Patterns in Signals, Speech, Images and Videos*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 306–312.
32. Peng, X.; Wu, X.; Peng, Q.; Qi, X.; Qiao, Y.; Liu, Y. Exploring dense trajectory feature and encoding methods for human interaction recognition. In Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service, ACM, New York, NY, USA, 17–19 August 2013; pp. 23–27.
33. Available online: https://www.semanticscholar.org/paper/The-Discrete-Cosine-Transform-(-DCT-)-%3A-Theory-and-Khayam/a99022c31741b797c609fde38286882e85a86d59 (accessed on 17 July 2022).
34. Cody, M.A. The fast wavelet transform beyond fourier transforms. *Dr. Dobb's J.* **1992**, *17*, 1–12.
35. Smith, J.; Chang, S.-F. Transform features for texture classification and discrimination in large image databases. In Proceedings of the 1st International Conference on Image Processing, Austin, TX, USA, 13–16 November 1994; Volume 3, pp. 407–411.
36. Climer, S.; Bhatia, S.K. Image database indexing using JPEG coefficients. *Pattern Recognit.* **2002**, *35*, 2479–2488. [CrossRef]
37. Gonzalez-Audicana, M.; Saleta, J.L.; Catalan, R.G.; Garcia, R. Fusion of multispectral and panchromatic images using improved ihs and pca mergers based on wavelet decomposition. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1291–1299. [CrossRef]
38. Mikolajczyk, K.; Schmid, C. Scale & affine invariant interest point detectors. *Int. J. Comput. Vis.* **2004**, *60*, 63–86.
39. Koornwinder, T.H. *Wavelets: An Elementary Treatment of Theory and Applications*; World Scientific: Singapore, 1993; Volume 1, ISBN 9789814503747.
40. Pearson, K. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [CrossRef]
41. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1993**, *24*, 417. [CrossRef]
42. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Meth-Ods*; Cambridge University Press: Cambridge, UK, 2000.
43. Mitchell, T. *Machine Learning*; McGraw Hill: New York, NY, USA, 1997; ISBN 0070428077.
44. Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **1999**, *10*, 988–999. [CrossRef]
45. Chang, C.-C.; Lin, C.-J. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27. [CrossRef]
46. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
47. Arunnehru, J.; Davi, A.K.N.; Sharan, R.R.; Nambiar, P.G. Human Pose Estimation and Activity Classification Using Machine Learning Approach. In *International Conference on Soft Computing and Signal Processing*; Springer: Singapore, 2019; pp. 113–123.
48. Arunnehru, J.; Kumar, A.; Verma, J.P. Early Prediction of Brain Tumor Classification Using Convolution Neural Networks. In *International Conference on Computational Intelligence, Security and Internet of Things*; Springer: Singapore, 2019; pp. 16–25.
49. Poonkodi, M.; Vadivu, G. Action recognition using correlation of temporal difference frame (ctdf)—an algorithmic approach. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *12*, 7107–7120. [CrossRef]
50. Ji, X.; Wang, C.; Ju, Z. A New Framework of Human Interaction Recognition Based on Multiple Stage Probability Fusion. *Appl. Sci.* **2017**, *7*, 567. [CrossRef]
51. Uddin, A.; Joolee, J.B.; Alam, A.; Lee, Y.-K. Human Action Recognition Using Adaptive Local Motion Descriptor in Spark. *IEEE Access* **2017**, *5*, 21157–21167. [CrossRef]
52. Liang, J.; Xu, C.; Feng, Z.; Ma, X. Affective interaction recognition using spatio-temporal features and context. *Comput. Vis. Image Underst.* **2016**, *144*, 155–165. [CrossRef]
53. Zhang, Y.; Liu, X.; Chang, M.-C.; Ge, W.; Chen, T. Spatio-Temporal Phrases for Activity Recognition. In *Computer Vision–ECCV 2012*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 707–721.
54. Vahdat, A.; Gao, B.; Ranjbar, M.; Mori, G. A discriminative key pose sequence model for recognizing human interactions. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 1729–1736.

55. Motiian, S.; Feng, K.; Bharthavarapu, H.; Sharlemin, S.; Doretto, G. Pairwise Kernels for Human Interaction Recognition. In *Advances in Visual Computing*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 210–221.
56. Ryoo, M.S.; Chen, C.-C.; Aggarwal, J.K.; Roy-Chowdhury, A. An Overview of Contest on Semantic Description of Human Activities (SDHA) 2010. In *Recognizing Patterns in Signals, Speech, Images and Videos*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 270–285.
57. Yao, A.; Gall, J.; Van Gool, L. A Hough transform-based voting framework for action recognition. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2061–2068.