A Novel Human Action Recognition and Behaviour Analysis Technique using SWFHOG

Aditi Jahagirdar[1]
Department of Information Technology
MIT College of Engineering, Savitribai Phule Pune
University, Pune, India

Manoj Nagmode[2]
Department of Electronics and Telecommunication
Government College of Engineering and Research
Avasari khurd, India

Abstract*—In this paper, a new local feature, called, Salient Wavelet Feature with Histogram of Oriented Gradients (SWFHOG) is introduced for human action recognition and behaviour analysis. In the proposed approach, regions having maximum information are selected based on their entropies. The SWF feature descriptor is formed by using the wavelet sub-bands obtained by applying wavelet decomposition to selected regions. To improve the accuracy further, the SWF feature vector is combined with the Histogram of Oriented Gradient global feature descriptor to form the SWFHOG feature descriptor. The proposed algorithm is evaluated using publicly available KTH, Weizmann, UT Interaction, and UCF Sports datasets for action recognition. The highest accuracy of 98.33% is achieved for the UT interaction dataset. The proposed SWFHOG feature descriptor is tested for behaviour analysis to identify the actions as normal or abnormal. The actions from SBU Kinect and UT Interaction dataset are divided into two sets as Normal Behaviour and Abnormal Behaviour. For the application of behaviour analysis, 95% recognition accuracy is achieved for the SBU Kinect dataset and 97% accuracy is obtained for the UT Interaction dataset. Robustness of the proposed SWFHOG algorithm is tested against Camera view angle change and imperfect actions using Weizmann robustness testing datasets. The proposed SWFHOG method shows promising results as compared to earlier methods.

Keywords—Action recognition; behaviour analysis; HOG; salient wavelet feature; neural network; wavelet transform; SWFHOG

I. Introduction

In the recent era, the ease of capturing videos with CCTV cameras and smartphones has increased the amount of available video data enormously. Analyzing this data manually has become a tedious and time-consuming task. Automatically recognizing the behaviour of a person as normal or abnormal, by detecting the action performed, can lead to more robust intelligent video surveillance system.

Automatic human action recognition plays an important role in many applications like intelligent video surveillance systems, Human-machine interaction, Health care, robotics, etc. As per the level of difficulty, actions are regarded as gestures, simple actions, interactions and, group activities. A gesture is a movement specifically done to give some meaningful message e.g. sign language. Simple actions are day to day activities like walking, running, jumping, etc., which can be considered as a sequence of gestures. In interactions, two humans or one human and one object are involved. Handshaking, hugging, a person lifting a bag, etc. can be considered as interactions. More than two people doing an action like talking, walking together, etc. are considered as a group activity. Various approaches have been proposed for recognizing all these types of actions. The Methodology used for human action recognition changes with the change in the complexity of action to be recognized.

Action recognition plays an important role in behaviour understanding tasks. Recognizing the action performed by a person can lead to the detection of abnormal behaviour or abnormal event like a fight between two people, a patient falling, etc. A behaviour understanding task can be considered as a human action recognition task where an action performed by a person is categorized as normal or abnormal. Most of the methods which used handcrafted features for representing the action used an approach shown in Fig. 1. It is having three main steps: feature extraction, dimensionality reduction, and pattern classification.

The main challenge in this approach is devising a robust feature vector that can tackle challenges like illumination changes, occlusion, camera jitter, etc. In this work, a new local feature, named Salient Wavelet Feature and Histogram of Oriented Gradients (SWFHOG) is introduced for the action recognition and behaviour analysis task. The feature is a combination of newly introduced Salient Wavelet Feature (SWF) and existing Histogram of Oriented Gradient (HOG) feature. To form the SWF feature, in the first step, salient regions are extracted by selecting areas of maximum motion and in the second step, average and detail wavelet coefficients are computed from these salient regions using the wavelet decomposition technique.

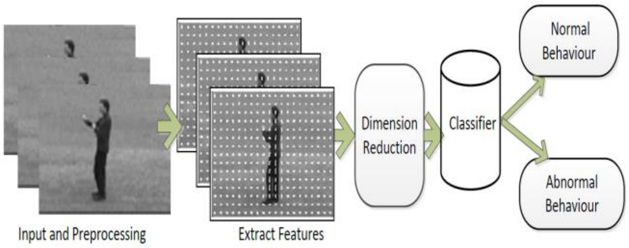


Fig. 1. General Human Action Recognition System for behavior Analysis.

II. Related Work

In this section, various methods proposed for human action recognition using handcrafted features are discussed. Features used in action classification are broadly divided as global features and local features. Global features describe the frame as a whole and generalize the object present in it. Local features treat a frame as a collection of small patches and describe them. Global features are useful in object detection, while the local features are more useful in object recognition. A combination of the global and a local feature is proved to increase the recognition accuracy of the system in most cases. Shape Matrices, Invariant Moments (Hu, Zernike), Histogram Oriented Gradients (HOG) and Co-HOG are some examples of global features used in action recognition. SIFT, SURF, LBP, BRISK, MSER and FREAK are some examples of local features used for action recognition [1] The emphasis of this related work is to review various methods that use salient point detection, wavelet transform as a feature and latest trends in action recognition.

Dawn et al. [2] have done the all-inclusive study of the use of Spatio Temporal Interest Point extraction methods in Human action recognition. Bak, Cagdas et al. [3] have proposed the use of saliency detection in videos for action recognition. Authors have used deep learning methods for saliency detection and various fusion mechanisms are studied for integrating spatial and temporal information. Ashwan Abdulmunem et al. [4] have proposed a method using salient object detection. The authors also propose a combination of a local and a global descriptor to classify the actions using the SVM classifier. Amir Ghodrati and Shohreh Ka-saei, in [5], have proposed methods for local spatiotemporal feature selection. The authors propose two weighing schemes to rank the features. Duta IC et al. [6] have proposed an extended version of the VLAD feature incorporating Spatial and Temporal information viz. ST-VLAD. The proposed method gives comparable results on datasets used for testing.

Al-Berry et al. [7], have proposed the use of Stationary Wavelet Transform (SWT) along with Local Binary Pattern (LBP) features to devise a feature descriptor. The proposed method achieves good accuracy on tested datasets. Al-Berry et al. [8, 9] and Siddiqi et al. [10] have used a combination of local and global features to construct a feature descriptor to take advantage of both the techniques. As wavelet coefficients represent multiscale and directional information of motion pattern, wavelet coefficients are used for describing the action. The use of a discrete wavelet transform for motion detection is explored by other researchers and proved to give good results [11-13]. As the number of interest points detected is large in number, many times they impose overhead on the further process. Some researchers have proposed approaches for extracting only important interest points before forming the feature descriptor. Bhaskar Chakraborty et al. [14, 15] have proposed a method to suppress the interest points from the background by maintaining only the repetitive and stable interest points. Bag of video words model, using N jet features is then applied for the representation of the action.

A detailed review of abnormal behaviour detection methods is given in [16, 17]. It is seen that analyzing the behaviour is nothing but recognizing the action performed by the person and then tagging the action with some behavioral name. The authors have shown that approaches like optical flow, STIP detection, HOG feature, Object tracking, and trajectory extractions, are used for behaviour analysis. In [18], a novel approach for behaviour recognition is proposed. The authors have proposed the use of a dynamic probabilistic graph for describing the temporal relationship between the objects. In [19], an approach based on pixel change history is proposed for behaviour analysis. The authors propose the use of two probabilistic masks one for face and another for body detection. HMM is used for recognition and classification.

From the literature review, it was observed that local features play an important role in discriminating between similar actions. The extraction of salient regions or objects from the video before extracting features increases the efficiency of the algorithm. In the existing methods, salient regions are selected based on response values computed at the pixels. These methods does not consider the salient regions as volumes and thus fail to detect volumes having maximum movement. The method proposed in this work uses the information content of 3D volume constructed around each interest point to select it as salient region. Wavelet coefficients of these salient regions are then extracted to form a local feature descriptor.

III. Proposed Algorithm using SWFHOG Feature

This section gives details of the proposed SWFHOG based human action recognition and behaviour analysis technique. Fig. 2 shows a block schematic of the proposed method. As shown in the diagram, SWF local feature and HOG global feature are computed for the video separately. Dimensionality reduction is achieved for the features by applying Principal Component Analysis (PCA). The two feature vectors thus obtained are combined to form a SWFHOG feature descriptor. Each block of the diagram is discussed in detail here.

A. *Input and Preprocessing*

The input to the system is action video clips. The input video is converted to frames and median filtering is applied to reduce the noise present in it. As each dataset is having different specifications, for ease of execution, all the frames are resized. A three-dimensional array of frames is formed and given as input to the next stage.

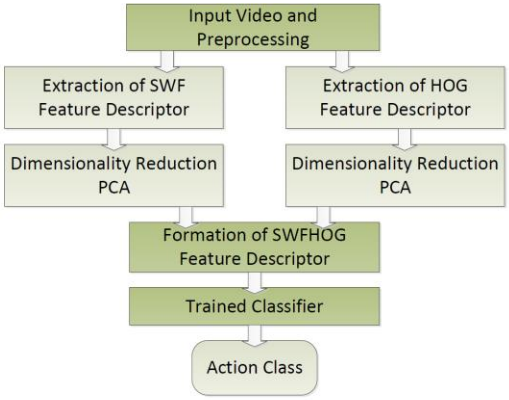

Input Video and Preprocessing
Extraction of SWF Feature Descriptor
Extraction of HOG Feature Descriptor
Dimensionality Reduction PCA
Dimensionality Reduction PCA
Formation of SWFHOG Feature Descriptor
Trained Classifier
Action Class

Fig. 2. Block Schematic of SWFHOG Feature Descriptor.

B. *Details of SWF Feature*

The proposed Salient Wavelet Feature is local. The main steps in SWF feature extraction are Salient region extraction and wavelet decomposition. In most of the action videos, the motion is present in a lesser amount of area of a frame as compared to the background area. In the videos where humans are present, significant motion is present in the region around the human figure. Such regions having maximum spatial and temporal changes are defined as regions of interest or salient regions.

In this work, for extracting the salient regions, interest points are identified using the method proposed by Dollar et al. in [20]. This method is having the advantage that it detects fewer interest points from the background as compared to those detected by methods proposed by Laptev and Lindeberg [21], and Willems et al. [22].

Here, a 2D Gaussian smoothing filter, as given in (1), is applied to each frame in the spatial domain.

$$g(x,y,\sigma) = \frac{1}{2\pi\sigma^2}\ e^{-(x^2+y^2)/2\sigma^2} \qquad (1)$$

The Gaussian filter is convolved with the frame in x and y-direction. The spatial variance σ^2 is used as a spatial scale in x and y-direction. A temporal filter is then applied in the t direction to the smoothed image. Here, two orthogonal 1D Gabor filters are used for temporal filtering. h_{ev} denotes the even part and h_{od} denotes the odd part of the filter. Squared product of the two 1D filters is computed to find the final response. Equations for Gabor filter is shown in (2).

$$h_{ev}(t,\tau,\omega) = -\cos(2\pi t\omega)\, e^{-t^2/\tau^2}$$

$$h_{od}(t,\tau,\omega) = -\sin(2\pi t\omega)\, e^{-t^2/\tau^2} \qquad (2)$$

The temporal variance τ^2 controls the temporal scale. Gabor filter is a linear filter and its direction and frequency response matches the human visual system. It is used mainly for edge detection in image processing applications. Gabor filter is also efficient in texture classification. These two properties of the Gabor filter make it a perfect candidate for interest point detection. The value of ω is selected to be 0.5 / τ as a correction factor. The intensity value at each pixel is then considered for identifying the interest points.

The response function R, which represents the intensity value at each pixel can be given as in (3). I represent the image intensity, $g(\sigma)$ represents the Gaussian function while $h_i(\omega,\tau)$ represents the Gabor function. Salient points are detected by finding the value of response function R at every point.

$$R = \sum_{i=1}^{2} (I * g(\sigma) * h_i(\omega,\tau))^2 \qquad (3)$$

Some of the interest points are detected from the background pixels. These are the false interest points and increase the overhead in further processing. Fig. 3 shows the different number of interest points selected for the sample frame of handshaking action video from the SBU Kinect dataset. In this video, motion is present in the regions of the joined hands of both the actors.

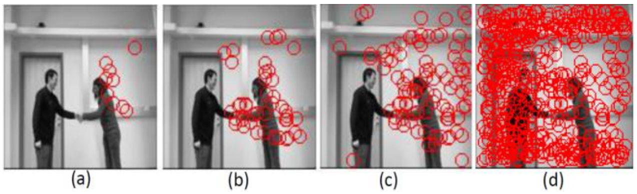


Fig. 3. The Number of Interest Points Selected. (a) k =10 (b) k=50 (c)k= 100 and (d) k= 500.

To remove the redundant interest points, first k significant interest points, having maximum response value, are selected. In the first iteration, a point having maximum response value is selected from the set of all the detected interest points and stored as a selected salient point in the subset (S). This point is then deleted from the set of all extracted interest points (L). In the next iteration, a point having maximum value is selected from the set of interest points having L-1 interest points. The process is repeated for the required number of times to extract the required number of interest points (k). It is seen that 10 points are not able to describe the movement in the action satisfactorily. For k=100 and k=500, many interest points are selected from the background. The interest points from the background do not contribute to describing the action. For k=50, the interest points selected are from the regions having maximum motion and are used in further processing.

After selecting the k salient points, a cuboid is extracted around each selected interest point by considering it as a center. The size of the cuboid in x and y direction depends on spatial scale σ while the size in z-direction depends on temporal scale τ. The cuboids thus extracted represent the regions of the video and are used in further process. In this work, the value of σ is selected as 2 whereas the value of τ is selected as 3. Fig. 4 is the visualization of sample cuboids of handshake video from the SBU Kinect dataset. Each row of in the diagram represents the journey of a small part of a frame through the temporal domain. The first row captures the movement of the hand of the actor. Ninth and tenth rows capture the movement of the head of two actors. Even after selecting the interest points with care, few of the cuboids carry information of the background pixels and do not contribute much to labeling the action.

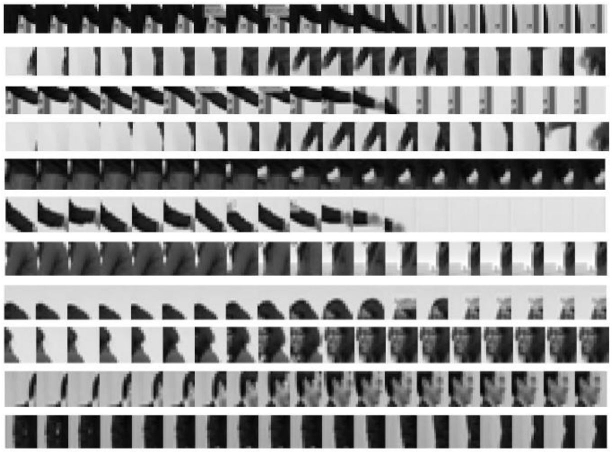


Fig. 4. Visualization of Sample Cuboids from Handshake Video.

To remove the cuboids having less information from further processing, the salient region extraction algorithm is used. The cuboids having maximum information are selected as salient regions. To find the information content, entropy is calculated for each cuboid. Entropy is a statistical measure used to find information present in an image. Entropy is calculated as given in (4).

$$H = -\sum_{i=1}^{k} P_k \, log_2 \, P_k \qquad (4)$$

Where H denotes the entropy and P_k denotes the probability associated with each grayscale in the image. Probability P_k is calculated by computing the histogram over all the gray scales.

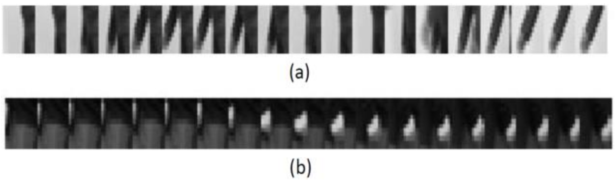
In this work, the number of cuboids extracted is equal to the number of interest points selected. If the number of interest points selected is k, spatial size is m x m and temporal size is n then, k cuboids of size m x m x n are formed. To compute the entropy of a cuboid, the entropy of each m x m part of the image is computed. Average entropy of n such m x m parts is computed for one cuboid and is stored as the entropy of that cuboid. The average of the entropies of all such k cuboids is then calculated and used as a threshold. The entropy of each cuboid is compared with the threshold value and cuboids having entropy more than the threshold are selected as Salient Regions. The steps of salient region extraction using the entropy of cuboids are shown in the algorithm here.

Algorithm: Salient Region Extraction

Input: Selected Interest Points (S)
Output: Selected salient regions

```
Begin
for $i = 1: s$
    $Cuboids_{All} = Form_{Cuboid}(\sigma, \square)$
end
for $i = 1: s$ ; where s is number of total cuboids
    for $j = 1: n$ ; where n is number of images in cuboid
        $Entropy_{im}(j) = Find_{Entropy}(im)$
    end
    $Cuboid_{entropy}(s) = Mean(Entropy_{im})$
end
$Threshold = Average(Cuboid_{entropy})$
for $i = 1: s$
    if $Cuboid_{entropy}(s) > Threshold$
        $Salient_{regions} = Cuboid_{entropy}$
    end
end
```

Fig. 5 shows the sample of (a) selected cuboid and (b) the rejected cuboid. The entropy of the cuboid in Fig. 5(a) is high (0.9324) as variation is present in it indicating the movement. The entropy of cuboid in Fig. 5(b) is very less (0.2642) as variation across the frames is very less indicating negligible movement. These selected cuboids are called as salient regions and are used in the further computation.


(a)

(b)

Fig. 5. (a) Selected and (b) Rejected Cuboid.

In the second step of the SWF algorithm, the wavelet decomposition technique is applied to the extracted salient regions. Average and detail coefficients are extracted from the salient regions to form a feature descriptor. While there are many types of wavelets, Daubechies wavelets (db) are most widely used because of their slightly longer support [23]. The db1 wavelet or Harr wavelet is used in this work as it is the simplest wavelet. The Haar wavelet is not differentiable as it is not a continuous function. This property of the Haar wavelet makes it useful for detecting sudden changes like motion present in action video. The steps to find the wavelet coefficients are given as:

1) Obtain low pass and high pass decomposition filter coefficients.

2) Convolve input image row-wise with low pass decomposition filter coefficients obtained in step 1.

3) Down-sample the output obtained in step 2 to keep only even indexed elements to get intermediate matrix z.

4) Convolve matrix z column-wise with low pass and high pass decomposition filter coefficients separately to obtain the average and detail horizontal coefficients.

5) Convolve input image row-wise with high pass decomposition filter coefficients obtained in step 1.

6) Down-sample to keep only even indexed elements to get intermediate matrix z.

7) Convolve matrix z obtained in step 3 column-wise with low pass and high pass decomposition filter coefficients separately to obtain detail vertical and detail diagonal coefficients.

The horizontal, diagonal and vertical coefficients are combined to form detail coefficients. The feature descriptor formed using average coefficients is named SWF_A whereas that formed using only detail coefficients is named SWF_D. Feature descriptor formed using average plus detail coefficients is called SWF_AD. Experimentation is done using all the three variants of the SWF.

C. Details of Histogram of Oriented Gradients Feature Descriptor

The proposed local SWF feature descriptor is combined with a Histogram of Oriented Gradients (HOG) global feature descriptor to form the SWFHOG feature descriptor. HOG has been proved to give good results for human action recognition and is explored by many researchers [24]. HOG feature descriptor represents the shape of an object within an image efficiently. As HOG was originally designed for person detection by Dalal and Triggs [25], it is a perfect candidate for human action recognition.

To find the HOG features, the image is divided into small patches called blocks (e.g. 16 x 16). Each block is further divided into cells (e.g. 8 x 8). 1-D centered, derivative masks are then applied in vertical and horizontal directions to compute gradients in x and y directions. [-1, 0, 1] and [-1, 0, 1]^T are proved to be good kernels for human detection. Gradients in x and y directions are computed as G_x and G_y respectively at each pixel, as given in (5), where, *I(x, y)* is the intensity at the pixel.

$$G_x(x,y) = I(x+1,y) - I(x-1,y)$$
$$G_y(x,y) = I(x,y+1) - I(x,y-1) \quad (5)$$

Magnitude G_{mag} and angle $G_{\emptyset}$ of the gradient at each pixel are then computed by using (6) and (7) respectively.

$$G_{mag}(\mathrm{x,y}) \; = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \quad (6)$$

$$G_{\emptyset} = arctan\frac{G_y(x,y)}{G_x(x,y)} \quad (7)$$

The histogram of the gradients is then formed for each cell. L2 normalization is then applied to each block to remove the effect of contrast variations. The final HOG feature consists of normalized histograms of each cell of each block of the image.

D. Dimensionality Reduction

The number of features extracted using the SWF algorithm as well as the HOG algorithm are large in number. Many of these features represent the background of the frame and contribute less to classification tasks. The features having less variance are redundant and can be removed from further processing. In this work, Principal Component Analysis is applied separately to SWF features and HOG features for achieving dimensionality reduction. Only the features having high variance are selected as final features.

E. Formation of SWFHOG Feature Descriptor

The SWF and HOG features obtained after applying dimensionality reduction are used in the construction of the SWFHOG feature descriptor. As shown in the results section, the performance of the SWF_AD feature is better than SWF_A and SWF_D features, for most of the datasets. This makes SWF_AD a perfect candidate for the SWFHOG feature descriptor. Both, SWF and HOG features are normalized to avoid the influence of any one feature on classification output. The concatenation of SWF_AD and HOG feature is done and is named as the SWFHOG feature descriptor.

SWF_AD local feature captures the motion information from the small patches of the video. Strong localization ability of Wavelet transform in spatial as well as frequency domain makes it possible to extract motion information in the form of wavelet coefficients from the video. Detail wavelet coefficients can capture minute movements happening in the small patches whereas average coefficients can describe the spatial information. The HOG feature is global and detects the shape of the human figure efficiently. In short, it can be said that, when the SWFHOG feature is extracted for an action video, HOG detects human silhouette from the frame whereas the SWF feature detects the movements of the body parts of the human. The selection of salient regions before applying wavelet decomposition makes it possible to reduce the redundancy and extract the local features having maximum information content. Thus the combination of SWF local feature and HOG global feature can describe the action efficiently.

F. Classifier

For classifying the actions using the proposed SWFHOG feature descriptor, a feed-forward neural network is used. The number of hidden layers used for good performance is determined empirically. For getting the unbiased estimate of the performance of the proposed descriptor, the dataset is divided into three parts namely, training data, testing data, and validation data. Random stratified sampling of the data is done. Data is repeatedly and randomly partitioned as training data and testing data in a predefined ratio. While randomly selecting the training and testing samples, it is ensured that class proportions are maintained as in the main dataset.

For all the experiments, 80% of samples are used for Training, 10% for validation and 10% for testing. Each set up is run 6 times considering different samples for Training, Validation, and Testing. Average Accuracy, Precision, Recall, and F1Score are then calculated.

IV. Experimental Results

Extensive testing is done to evaluate the performance of the proposed SWFHOG feature descriptor. Three experimentation setups are run for evaluating the proposed algorithm. In the first set up, the use of wavelet coefficients for the action recognition task is explored by using different groups of average and detail sub-bands. Accuracy and F1Score are computed for each action class. Overall accuracy and F1Score are then computed by taking the average of values obtained for all the classes. In the second set up, the use of the proposed algorithm for behaviour analysis is studied. An event can be labeled as Normal or Abnormal depending on the behaviour pattern identified. The actions of UT Interaction and SBU Kinect dataset are divided into two sets as Normal behaviour and Abnormal Behaviour for this experimentation. In the third set up, the robustness of the proposed algorithm against imperfect actions and camera view angle change is tested. This section discusses the datasets used for testing and the results obtained with the proposed SWFHOG feature descriptor.

A. Datasets used

This section gives brief information about the datasets used for testing the proposed algorithm. Weizmann, KTH, UCF Sports and UT interaction action datasets are used for evaluating the performance of the proposed method for action recognition. SBU Kinect Two-Person Interaction dataset and UT Interaction dataset are used for behaviour analysis. To evaluate the robustness of the proposed method against imperfect actions and camera view angle change, Weizmann robustness testing and Weizmann view angle change datasets are used.

The Weizmann [26] and KTH [27] datasets have simple actions like running, walking, jogging, etc. recorded in a controlled environment. Videos in both these datasets have low resolution making it challenging. In the KTH dataset one action is recorded in four different scenarios like indoor, outdoor, with different types of cloths and with a different scale. This adds to the complexity of the dataset. UCF Sports dataset [28, 29] has video clips recorded at various sports events and is a realistic dataset. Cluttered backgrounds, different camera view angles, different scales, illumination changes and multiple people present in one frame are the complexities present in this dataset. Along with these complexities, high intra-class variation present in this dataset makes it a challenging dataset.

UT Interaction dataset [30] and SBU Kinect Two-person Interaction dataset [31] have the videos of interactions between two people. The actions handshaking, hugging, pointing a finger and approaching a person are considered as Normal behaviour. The actions push, punch and kick are considered as Abnormal behaviour.

Weizmann robustness testing and camera view angle change dataset are specifically recorded with some challenges. Weizmann robustness testing dataset is having videos in three categories. It has actor walking in unusual way, actor walking with an object and partially occluded action.

The Weizmann camera view angle change dataset is having a videos of a walking action recorded with ten different camera view angles ranging from 0^0 to 90^0. Both these datasets are recorded in a realistic environment and have a cluttered background. Fig. 6 shows sample frames from all the datasets used.

B. Performance Parameters used

To evaluate the performance of the proposed algorithm, Recognition accuracy and F1Score are used as performance parameters. These parameters are computed using True Positive, True Negative, False Positive and False Negative predicted values.

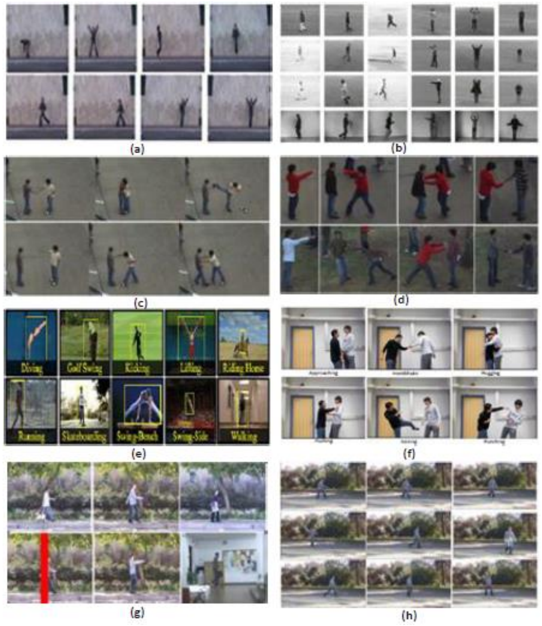


Fig. 6. Sample Frames from Action Datasets (a) Weizmann, (b) KTH, (c) UT1, (d) UT2, (e) UCF Sports (f) SBU Kinect Interaction (g) Weizmann Robustness Testing and (h) Weizmann Camera view Angle Change Dataset.

Recognition accuracy gives the ratio of correctly detected samples to the total number of samples. Precision and Recall becomes more important parameters in some action recognition applications. As precision and recall are inversely proportional to each other, to achieve the balance between these two metrics, the harmonic mean of precision and recall, called F1Score is calculated.

C. Experimental Setup 1

In this setup, the performance of different SWF variants is compared. Feature descriptor SWF_A, SWF_D, and SWF_AD are formed using only average coefficients, only detail coefficients and both the coefficients respectively. Performance is also compared with that achieved by the SWFHOG feature descriptor.

Detail analysis of results obtained for all the datasets is done. Table I illustrates the detail results obtained on the UT interaction1 dataset for intermediary execution. It gives action classification accuracy, precision, recall, and F1Score calculated from values of TP, TN, FP, and FN. Class 1 to class 6 represent actions punch, kick, hug, point a finger, handshake and push respectively.

TABLE I. DETAIL PERFORMANCE ANALYSIS ON UT INTERACTION DATASET

Algorithm	Class	TP	TN	FP	FN	Recall	Precision	Accuracy	F1Score
SWF_A	1	10	50	0	0	100	100	100	1
	2	9	50	0	1	90	100	98.33	0.95
	3	9	50	0	1	90	100	98.33	0.95
	4	7	49	1	3	70	87.5	93.33	0.78
	5	9	46	4	1	90	69.23	91.67	0.78
	6	10	49	1	0	100	90.91	98.33	0.95
SWF_D	1	8	50	0	2	80	100	96.67	0.89
	2	10	50	0	0	100	100	100	1
	3	10	50	0	0	100	100	100	1
	4	7	50	0	3	70	100	95	0.82
	5	9	45	3	1	90	75	93.1	0.82
	6	10	49	1	0	100	90.91	98.33	0.95
SWF_AD	1	10	50	0	0	100	100	100	1
	2	10	50	0	0	100	100	100	1
	3	10	50	0	0	100	100	100	1
	4	7	47	0	3	70	100	94.74	0.82
	5	9	46	3	1	90	75	93.22	0.82
	6	10	49	1	0	100	90.91	98.33	0.95
SWFHOG	1	10	50	0	0	100	100	100	1
	2	8	50	0	2	80	100	96.67	0.89
	3	10	50	0	0	100	100	100	1
	4	10	50	0	0	100	100	100	1
	5	9	51	0	1	90	100	98.36	0.95
	6	10	49	1	0	100	90.91	98.33	0.95

Table I shows that, for the SWF_A algorithm, more than 90% recognition accuracy is achieved for all the classes but less F1Score is obtained for classes 4 and 5. This is because of the lower values obtained for recall and precision. For the SWF_D algorithm, recognition accuracy gained is more than that in the case of SWF_A for all six classes. F1Score for classes 4 and 5 is improved than in the previous case but reduced for class 1. Since the SWF_AD algorithm gives high accuracy and F1score values for most of the cases, it is used to fuse with the HOG feature to form the SWFHOG feature. As seen from Table I, for the SWFHOG algorithm, high values of recall and precision are achieved for all the classes.

The graph in Fig. 7 shows the comparison of average recognition accuracies achieved with SWF_A, SWF_D, SWF_AD and SWFHOG feature vectors for all the datasets. The recognition accuracy values mentioned are computed by taking the average of classification accuracy values obtained for all the action classes after running the program multiple times. Table II shows the values obtained.

It is seen that higher recognition accuracy is obtained by the SWF_AD feature as compared to that obtained by SWF_A and SWF_D features individually, for all the datasets except the KTH dataset. As average wavelet coefficients capture low-frequency information while detail coefficients capture high-frequency information, their combination tends to give better results as compared to individual coefficients. The last row of Table II gives recognition accuracy obtained with the proposed SWFHOG feature descriptor. The highest recognition accuracy is obtained with the SWFHOG descriptor as compared to other variants.

The proposed feature descriptor is also evaluated based on F1Score to take into account the effect of all the SWF variants on precision and recall values.

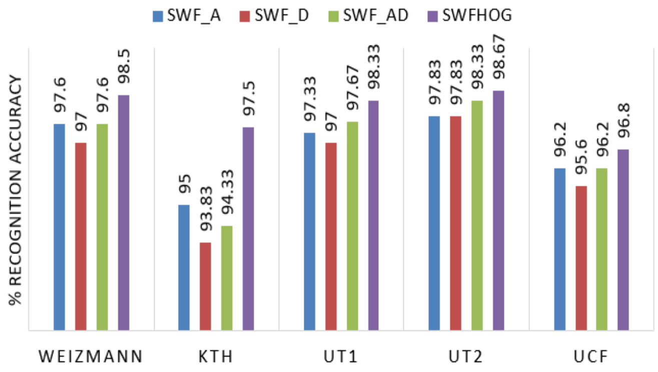


Fig. 7. Comparison of % Recognition Accuracy Obtained for Action Recognition.

TABLE II. % RECOGNITION ACCURACY ACHIEVED WITH SWF VARIANTS

	% Recognition Accuracy				
	Weizmann	**KTH**	**UT1**	**UT2**	**UCF**
SWF_A	97.6	95	97.33	97.83	96.2
SWF_D	97	93.83	97.00	97.83	95.6
SWF_AD	97.6	94.33	97.67	98.33	96.2
SWFHOG	98.5	97.5	98.33	98.67	96.8

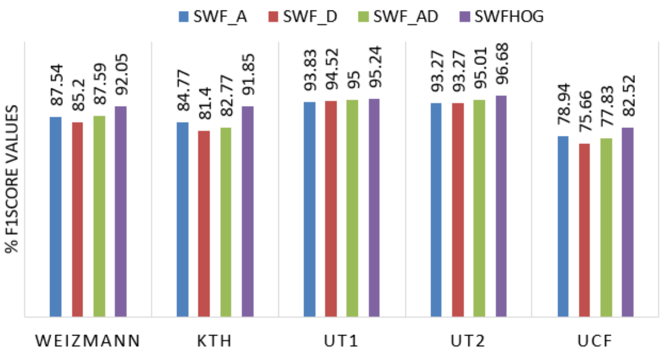


Fig. 8. Comparison of % F1 Score Obtained for Action Recognition.

The graph in Fig. 8 shows the F1Score values achieved for all the datasets using variants of the SWF feature. Table III gives the values of the F1Score obtained. It is seen that high values of the F1Score are obtained for all the datasets when the SWFHOG feature is used. The SWFHOG algorithm can represent each action most distinctly, reducing false positive and false negative classifications. This results in the increase in the values of precision and recall which reflects in the escalation in the F1Score value.

TABLE III. % F1SCORE ACHIEVED WITH SWF VARIANTS

	%F1Score values				
	Weizmann	**KTH**	**UT1**	**UT2**	**UCF**
SWF_A	87.54	84.77	93.83	93.27	78.94
SWF_D	85.2	81.4	94.52	93.27	75.66
SWF_AD	87.59	82.77	95	95.01	77.83
SWFHOG	92.05	91.85	95.24	96.68	82.52

D. Experimental Setup 2

The proposed SWFHOG algorithm is evaluated for the behaviour analysis. When two people interact, the action performed can be friendly, like a handshake, or can be unfriendly, as a person pushing the other. In this work, the behaviour is discriminated against as Normal and Abnormal.

The action videos from UT Interaction 1, UT Interaction 2 and SBU Kinect two-person interaction dataset are divided into two categories as Normal and Abnormal. For the UT Interaction dataset, actions, "Handshake", Hug" and "Point a Finger" as a normal action. Actions "Push", "Punch" and "Kick" are considered abnormal actions. For the SBU Kinect dataset, only RGB data is used in this experimentation. Interactions, "Person Approaching", "Hugging" and "Handshaking" are considered as normal behaviour whereas actions "Kicking", "Pushing" and "Punching" are considered as abnormal behaviour. Binary classification is performed using these two sets of videos.

It is very important to identify any abnormal event as abnormal in the case of video surveillance. Only recognition accuracy is not sufficient to decide about the performance of the classification algorithm in this case. The recall is a parameter that tells how many samples are detected correctly as compared to the actual true samples. This means that true positive detections should be maximized and false negative values should be minimized. This means that the high value of Recall is desirable. To take into account this fact, recall values are also computed for all datasets. Table IV shows the results obtained.

Results show that more than 97% recognition accuracy, as well as recall value, is obtained for the UT Interaction dataset. For the SBU Interaction dataset, more than 95% recognition accuracy, as well as the recall, is achieved. In the behaviours which are considered normal (handshake, hug, approach) in this setup, two people approach each other and then stay in the same position. In the actions which are considered abnormal (push, punch, kick), two people approach each other and move back from each other at the end of the action. The proposed SWFHOG feature can distinguish between these two patterns satisfactorily.

E. *Experimental Setup 3*

To evaluate the robustness of the proposed SWFHOG algorithm to high regularities like occlusion, unusual way of performing the action, varied background and view angles, Weizmann robustness dataset is used for testing. Table V shows the recognition accuracy obtained for the robustness testing dataset. It is observed that the average recognition accuracy of more than 94% is achieved for the Weizmann robustness testing action dataset. The proposed SWFHOG algorithm can recognize the action as walking cation 18 times out of 19. It was seen that, as the view angle approaches, 90^0 (Person approaching camera), action recognition becomes more difficult as scale in the sequence changes substantially. The proposed SWFHOG algorithm can recognize the walk action correctly even if the clothing of actors is different, actors are walking unusually or are walking with a bag in hand. This proves the robustness of the proposed SWFHOG method.

TABLE IV. PERFORMANCE OF SWFHOG FOR BEHAVIOUR ANALYSIS

Parameter Dataset	% Recognition Accuracy	% Recall
UT Interaction 1	97.08	97.19
UT Interaction 2	97.92	97.92
SBU Kinect Interaction	95.74	95.92

TABLE V. PERFORMANCE ON THE WEIZMANN ROBUSTNESS TESTING DATASET

S. No.	Details of the samples used	%Recognition Accuracy
1	9 samples of Normal walk + 10 samples of unusual walk	94.70
2	90 samples of 10 classes + 10 samples of unusual walk	93.60
3	90 samples of 10 classes + 10 samples of the walk with view angle change	94.82
4	81 samples of 9 classes (other than normal walk) + 10 samples of unusual walk	94.55
5	81 samples of 9 classes (other than normal walk) + 10 samples of the view angle change	94.49

F. *Comparison of the Proposed Method with Existing Methods*

Table VI gives a comparison of recognition accuracy achieved for the UCF Sports and UT Interaction dataset by the proposed SWFHOG method and the existing methods. Methods that have used handcrafted features are used for comparison. Accuracy values mentioned in the table are taken from papers published by various researchers. It is observed that the SWFHOG method gives average recognition accuracy of 96.8% for the UCF sports dataset which is higher than other existing methods. For the UT interaction dataset, recognition accuracy outperforms all the existing methods with a recognition accuracy of 98.5% (calculated by taking an average of values obtained for UT Interaction 1 and UT Interaction 2).

Table VII gives a comparison of recognition accuracy achieved for the KTH and Weizmann dataset by the proposed SWFHOG method and the existing methods. The comparison shows that the performance achieved by the SWF_H method outperformance most of the existing methods. For the Weizmann dataset, slightly higher accuracy is achieved with a structural average based method [20]. For the KTH dataset, a method based on Log-Euclidean covariance matrices of ST features [17] achieves accuracy comparable with that achieved with the proposed method.

TABLE VI. COMPARISON: UCF AND UT INTERACTION DATASETS

Existing and Proposed methods	% Recognition Accuracy	
	UCF Sports	UT Interaction
Motion and appearance Saliency and trajectories [37]	90	--
Edge trajectories and Motion descriptor [38]	92	--
Edge trajectories and Spatiotemporal motion skeleton [39]	92.8	--
Temporal trajectories [40]	--	91.8
The BoW of interest points and HOG [41]	--	83.3
Key poses [42]	--	85%
SWFHOG (Proposed method)	**96.8**	**98.33**

TABLE VII. COMPARISON: KTH AND WEIZMANN DATASETS

Existing and Proposed methods	% Recognition Accuracy	
	KTH	Weizmann
Riemannian manifolds [32]	--	96.7
log-Euclidean covariance matrices of ST features [33]	97.1	---
A mixture of features [34]	92.28	91.69
Optical flow-based [35]	94.62	---
Structural average curves analysis [36]	---	98.77
SWFHOG (Proposed method)	**97.5**	**98.5**

V. Conclusion

In this paper, a new local feature, SWF, is introduced for representing human actions. Experimentation is done using combinations of sub-bands obtained from wavelet decomposition. To improve the performance further, SWF is used along with the HOG feature, which creates a robust combination of a local and global feature. Experimental results show that new local feature descriptor SWF, captures local features efficiently and when combined with HOG, outdoes accuracy achieved by most of the existing methods for UT interaction and UCF sports datasets. The proposed SWFHOG feature descriptor achieves good accuracy for Weizmann and KTH datasets.

Extracting the Salient regions increases the classification accuracy of the algorithm as only the cuboids having maximum information are used to form the descriptor. Strong localization ability of Wavelet transform in spatial as well as frequency domain makes it possible to extract motion information in the form of wavelet coefficients from the video. SWFHOG feature becomes robust against illumination changes because of the block normalization used while extracting the HOG feature. The proposed approach eliminates the requirement of the crucial task of segmentation and foreground extraction. The 94.55% accuracy obtained for imperfect action sequences and 94.49% accuracy achieved for sequences recorded with varied camera view angle prove the robustness of this algorithm. 97.92% accuracy and recall values achieved for UT interaction 2 dataset 95.72% and 95.92% accuracy and recall are achieved respectively for behaviour analysis. These results indicate the usefulness of proposed method for behaviour analysis.

Comparison of the results obtained by proposed algorithm with existing methods show that, the proposed SWFHOG method outperforms existing methods for UT Interaction and UCF Sports dataset. Recognition accuracy of 98.33% and of 96.8% is achieved for these two datasets for action recognition task. The SWFHOG algorithm gives high F1Score values, indicating that precision and recall values are well balanced.

The results in three experimental setups indicate that the SWFHOG feature algorithm combines advantages of global feature and local features, producing a strong feature descriptor for action recognition as well as behaviour analysis.

VI. Future Scope

In this work an approach for human action recognition based on new local feature descriptor is proposed. The proposed SWFHOG method is tested for recognizing a single action performed by an individual or a pair of individuals. In future, method can be devised to recognize multiple actions present in one video. The real world videos multiple humans performing various actions present in one video. Also recognizing multiple actions performed by a single person in one video remains a challenging task.

References

[1] Z. Hong-Bo, Y. Zhang, B. Zhong, Q. Yang, J. Du, and D. Chen, "A comprehensive survey of vision-based human action recognition methods," Sensors 19, no. 5, 1005, 2019.

[2] D. Dawn, D. Das, and S. Shaikh, "A comprehensive survey of human action recognition with Spatio-temporal interest point (STIP) detector," The Visual Computer Vol. 32, no. 3, pp 289-306, 2016.

[3] C. Bak, K. Aysun k, E. Erkut, and E. Aykut, "Spatio-temporal saliency networks for dynamic saliency prediction," IEEE Transactions on Multimedia, Vol. 20, no. 7, pp 1688-1698, 2017.

[4] A. Abdulmunem, L. Yu-Kun, and S. Xianfang, "Saliency guided local and global descriptors for effective action recognition," Computational Visual Media, Vol. 2, no. 1, pp 97-106, 2016.

[5] A. Ghodrati, and K. Shohreh, "Human action categorization using discriminative local Spatio-temporal feature weighting," Intelligent Data Analysis, Vol.16, no. 4, pp 537-550, 2012.

[6] I. Duta, I. Bogdan, A. Kiyoharu and S. Nicu, "Spatio-temporal vlad encoding for human action recognition in videos," In International Conference on Multimedia Modeling, Springer, Cham, pp. 365-378, 2017.

[7] M. Al-Berry, Mohammed A-M. Salem, et al., "Action Classification Using Weighted Directional Wavelet LBP Histograms," In The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), Beni Suef, Egypt, Springer, Cham. pp. 15-24, 2016.

[8] M. Al-Berry, Mohammed A-M. Salem, M. Hala, H. Ashraf, and M. Tolba, "Weighted Directional 3D Stationary Wavelet based Action Classification," Egyptian Computer Science Journal, Vol 39, no. 2, pp 83-97, 2015.

[9] M. Al-Berry, H. Ebied, A. Hussein and M. Tolba, "Human action recognition via multi-scale 3D stationary wavelet analysis," 14th International Conference on Hybrid Intelligent Systems, Kuwait, pp. 254-259, 2014.

[10] M. Siddiqi, M. Rana, E. Hong, K. Eun, and S. Lee, "Video-based human activity recognition using multilevel wavelet decomposition and stepwise linear discriminant analysis," Sensors, Vol. 14, no. 4, pp 6370-6392, 2014.

[11] S. Yousefi, MT Shalmani, J. Lin, M. Staring, “A Novel Motion Detection Method Resistant to Severe Illumination Changes”, arXiv preprint arXiv:1612.03382, December 2016.

[12] C. Hsia, J. Chiang, J. Guo, H. Olkkonen, “Multiple moving objects detection and tracking using discrete wavelet transform”, In Discrete Wavelet Transforms-Biomedical Applications, pp 297-320, Sep 12, 2011.

[13] S. Yousefi, M. Shalmani, J. Lin, M. Staring, “A Novel Motion Detection Method Using 3D Discrete Wavelet Transform”, IEEE Transactions on Circuits and Systems for Video Technology. December 5, 2018.

[14] B. Chakraborty, M. Holte, T. Moeslund, J. Gonzàlez, “Selective Spatio-temporal interest points” Computer Vision and Image Understanding. Vol.116, No. 3, pp 96-410, March 2012.

[15] B. Chakraborty, M. Holte, T. Moeslund, J. Gonzàlez, F. Roca, “A selective spatio-temporal interest point detector for human action recognition in complex scenes”, In International Conference on Computer Vision, pp. 1776-1783, IEEE, Nov 2011.

[16] A. Mabrouk, E. Zagrouba, “Abnormal behavior recognition for intelligent video surveillance systems: A review”, Expert Systems with Applications, Vol. Jan 1:91, pp 480-91, January 2018.

[17] O. Popoola, K. Wang, “Video-based abnormal human behavior recognition—A review”, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Jan 12;42(6), pp 865-78, 2012.

[18] S. Gong, "Towards behaviour recognition based video surveillance", In Optics and Photonics for Counterterrorism and Crime Fighting, vol. 5616, pp. 1-15. International Society for Optics and Photonics, 2004.

[19] A. Voulodimos, N. Doulamis, and S. Tafarakis, "Behavior recognition from video based on human constrained descriptor and adaptable neural networks", In Proceedings of the 4th ACM/IEEE international workshop on Analysis and retrieval of tracked events and motion in imagery stream, pp. 59-66. 2013.

[20] P. Dollár, R. Vincent, C. Garrison and B. Serge, "Behavior recognition via sparse Spatio-temporal features," Beijing, China: VS-PETS, 2005.

[21] I. Laptev and T. Lindeberg, " Space-time interest points", in ICCV, 2003.

[22] G. Willems, T. Tuytelaars, L. Van Goo, "An efficient dense and scale-invariant spatio-temporal interest point detector", In European conference on computer vision, pp. 650-663, Springer, Berlin, Heidelberg, 2008 Oct 2012.

[23] J. Walker, "A primer on wavelets and their scientific applications," CRC Press, 2002.

[24] A. Jahagirdar, M.Nagmode, "Human Action Recognition using Ensemble of Shape, Texture and Motion features", International Journal of Pure and Applied Mathematics, Vol. 119, No. 12, 2018, pp13025-13033.

[25] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," In Proc. of IEEE Computer Society Conference Computer Vision and Pattern Recognition, Vol. 1, pp. 886-893, 2005.

[26] E. Shechtman, L. Gorelick, M. Blank, M. Irani and R. Basri, "Actions as space-time shapes," IEEE Trans. Pattern Analysis and Machine Intelligence, 29(12), pp.2247-2253, 2007.

[27] C. Schuldt, I. Laptev and B. Caputo, "August. Recognizing human actions: a local SVM approach," In Proceedings of the 17th International Conference on Pattern Recognition, Vol. 3, pp. 32-36, IEEE, 2004.

[28] D. Mikel, J. Rodriguezd and M. Shah, "Action MACH: A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition," Computer Vision and Pattern Recognition, pp 1-8, 2008.

[29] K. Soomro and Amir R. Zamir, Action Recognition in Realistic Sports Videos, Computer Vision in Sports. Springer International Publishing, pp 181-208, 2014.

[30] M. Ryoo, C. Chen, J. Aggarwal and A. Roy-Chowdhury, "An overview of the contest on semantic description of human activities (SDHA)", In International Conference on Pattern Recognition, Springer, Berlin, Heidelberg, pp. 270-285, August 2010.

[31] K. Yun, J. Honorio, D. Chattopadhyay, T. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 28-35, 2012.

[32] J. Carvajal, W. Arnold et al. "Comparative evaluation of action recognition methods via Riemannian manifolds, Fisher vectors, and GMMs: Ideal and challenging conditions," In Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Cham, pp. 88-100, 2016.

[33] S. Cheng, J. Yang, Z. Ma, and M. Xie, "Action recognition based on Spatio-temporal log-Euclidean covariance matrix," International Journal of Signal Processing, Image Processing, and Pattern Recognition, Vol. 9, no. 2, pp 95-106, 2016.

[34] N. Humza, G. Khan, A. Khan, A. Siddiqi, and M. Ghani Khan, "Human activity recognition using a mixture of heterogeneous features and sequential minimal optimization," International Journal of Machine Learning and Cybernetics, Vol.10, no. 9, pp 2329-2340, 2019.

[35] S. Kumar and M. John, "Human activity recognition using optical flow-based feature set," In 2016 IEEE international Carnahan conference on security technology (ICCST), pp. 1-5, 2016.

[36] S. Zeng, G. Lu, and P. Yan, "Enhancing human action recognition via structural average curves analysis," Signal, Image and Video Processing, Vol.12, no. 8, 1551-1558, 2018.

[37] Y. Yi, and Y.Lin. "Human action recognition with salient trajectories," Signal Processing, Vol. 93, no. 11, pp 2932-2941, 2013.

[38] X. Wang, and Q. Chun, "Action recognition using edge trajectories and motion acceleration descriptor," Machine Vision and Applications, Vol 27, no. 6, pp 861-875, 2016.

[39] Z. Weng and G. Yepeng, "Action recognition using length-variable edge trajectory and spatio-temporal motion skeleton descriptor," EURASIP Journal on Image and Video Processing, no. 1, p 8, 2018.

[40] M. Saeid, F. Siyahjani, R. Almohsen, and G. Doretto, "Online human interaction detection and recognition with multiple cameras," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 27, no. 3, pp 649-663, 2016.

[41] J. Xiaofei, C. Wang, X. Zuo, and Y. Wang, "Multiple feature voting based human interaction recognition," International Journal of Signal Processing, Image Processing, and Pattern Recognition, Vol.9, no. 1, pp 323-334, 2016.

[42] S. Yasaman, A. Vahdat, S. Se, and G. Mori, "Discriminative key-component models for interaction detection and recognition," Computer Vision and Image Understanding, Vol.135, pp 16-30, 2015.