SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
Volume 6 No. 6
ISSN 2054-7390

Video Based Human Activity Detection, Recognition and Classification of actions using SVM

[1]Jagadeesh B, [2]Chandrashekar M Patil
[1]Research Scholar, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India;
[2]Professor, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India;
jagadeesh.b@vvce.ac.in; patilcm@vvce.ac.in

ABSTRACT

Human motion analysis which includes activity detection and action recognition is currently gaining attention from computer vision researchers. Automatic monitoring of human activities and actions using computers has found significant applications in video surveillance, monitoring of patients and sports applications. With the tremendous advancement and development in digital video libraries, automatic interpretation of videos will save human effort in analysis and interpretation. This has led to the development of robust techniques in the field of computer vision. Human activity detection and recognition includes detection of human, tracking of human and recognition of actions. In this paper, detection of human is done using Gaussian Mixture Model, tracking is done using optical flow, recognition and classification of actions is done using SVM Classifier. The experiment is carried out with two public datasets KTH and Weizmann which are the videos with constant background. The classification accuracy for KTH dataset is 92.48% and for Weizmann dataset the classification accuracy is 93.64%.

Keywords: Action Recognition; Human Motion Analysis; Video Surveillance; Gaussian Mixture Model; SVM Classifier.

1 Introduction

Human activity detection and recognition is a significant area in the field of computer vision research. The main aim of human activity recognition is to automatically analyze the activities from the videos and classify the videos appropriately into its activity category.

There are several categories of human activities. Contingent upon its intricacy, the human activities can be arranged into four levels: gestures, actions, interactions and group activities. Gestures are basic movements of parts of a human body, actions are individual person activities, interactions are activities involving two or more persons and group activities are the activities performed by abstract groups comprising of multiple persons and objects. Currently, daily activities and surveillance videos has provided rich video content and there is a problem in categorizing the videos based on the action class. Classifying the videos consumes a lot of time if it is manually done with large video database. Therefore there is a need to develop a robust automated action recognition system. Most of the approaches aim at solving the action recognition problem using pattern recognition system. In these approaches, feature descriptors from the videos are extracted and trained for each action video. This enables the digital computer to identify the actions in the videos automatically.

DOI: 10.14738/tmlai.66.5287
Publication Date: 25th November, 2018
URL: http://dx.doi.org/10.14738/tmlai.66.5287

Action recognition in videos still remains challenging due to variation in viewpoint, illumination, presence of occlusions and background changes. Realistic videos possess wide differences in posture, viewpoint, background, video resolution and occlusions. Also dealing with camera motion, zooming and non-static background is a serious issue to be addressed for a robust action recognition system. In this direction, wide research is being carried out.

The key contributions of this paper are threefold: Initially, for detection and tracking of human, GMM and optical flow methods are used. Next, for effective representation of features, Scale Invariant Feature Transform (SIFT) is used. Finally, in order to classify the actions, SVM based classification model is constructed.

The rest of the paper is structured as follows: Section 2 presents the related work carried in the field of action recognition, section 3 gives the detailed proposed framework, section 4 provides the results and discussion of the proposed work and section 5 briefs about the conclusion and future work.

2 Existing Related Work

Dadi, H.S., Pillutla, G.K.M. & Makkena, M.L. [1] presented a novel algorithm for face recognition and human tracking. Human is tracked using Gaussian Mixture Model. To track the human in precise, template of GMM is divided into four regions which are placed one above the other and tracked simultaneously. For recognizing the human, the Histogram of Oriented Gradients (HOG) features of the face region are given to the support vector machine classifier. Three experiments are conducted in taking the training faces. Every 10th frame, every 5th frame and every 3rd frame of the first 100 frames are considered. The other frames in the video are considered for testing using SVM classifier. Three datasets namely AITAM1 (simple), AITAM2 (moderate) and AITAM3 (complex) are used in this work. This is experimented for all types of datasets. The Performance results show that the combination of the tracking algorithm and the face recognition algorithm not only tracks the person but also recognizes the person [2].

Kiran Kale, Sushant Pawar, Pravin Dhulekar [3] developed novel object detection and tracking algorithm which uses optical flow in combination with motion vector approximation for object detection and tracking in a sequence of frames. The optical flow provides information about the object movement even if no computable parameters are computed. The motion vector estimation technique can provide an estimation of object position from consecutive frames which increases the accuracy of this algorithm and helps to provide robust result irrespective of image blur and cluttered background. The use of median filter with this algorithm makes it more robust in the presence of noise [4].

Ruichen Jin, Jongweon Kim [5] presented a scheme for tracking of object movements and detecting of feature to find video content by means of improved Scale-Invariant Feature Transform (SIFT). SIFT can robustly find objects even in clutter and under partial occlusion, because the SIFT feature descriptor is invariant to uniform scaling, orientation, and also partially invariant to affine distortion and illumination changes. Even if the video drops frames or attacked, this method can extract the features. In this method, video features from tracking the object's movement are detected and a dataset with feature sequences is used to identify video. In contrast to the existing tracking techniques, this method recognized reliable object coordinate. The developed algorithm will be an essential part of a completely tracking and identification system. To evaluate the performance of the proposed approach, experimentation was done with several genres of video.

Mona M.Moussa, Elsayed Hamayed ,Magda B.Fayek, Heba A.El Nemr [6] presented a fast and simple method for human action recognition. The proposed technique relies on detecting interest points using SIFT (scale invariant feature transform) from each frame of the video. A fine-tuning step is used here to limit the number of interesting points according to the amount of details. Then the popular approach Bag of Video Words is applied with a new normalization technique. This normalization technique remarkably improves the results. Finally a multi class linear Support Vector Machine (SVM) is utilized for classification.

Dhulavvagol P.M., Kundur N.C. [7] proposed a combination of two different techniques i.e. SVM and SIFT techniques to identify and recognize the human actions in a given video or image. To extract local features of the given video SIFT based technique is used. In this method, initially features based on the interest points at a particular point or frames are extracted. Once the key features are extracted they are further classified using SVM classifier [8].

3 Proposed Framework

Many approaches have been proposed to address the problem of action recognition. In this paper, the issue of detection of human, tracking, feature extraction and classification of action in videos is discussed. The proposed framework is as shown in the figure 1. The input video sequences are the public video datasets KTH and Weizmann dataset.

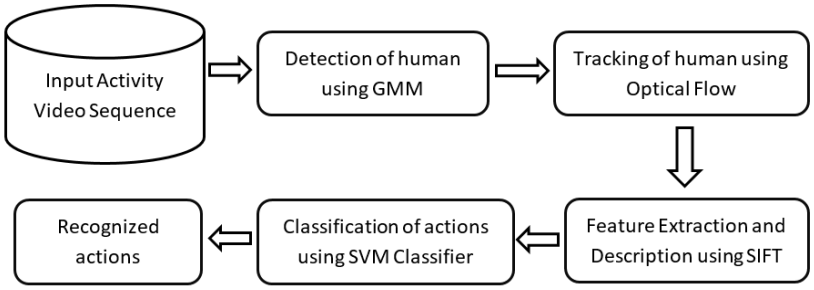


Figure 1: Proposed Framework of action recognition system

3.1 Detection of Human using GMM

Gaussian Mixture Model (GMM) is essentially one of the most well-known procedures to construct the background model for segmentation of moving objects from background. GMM technique assigns number of Gaussian distributions for each pixel to estimate reference frame. If there are no any variations in the pixel values then all Gaussian distributions approximate the same values. In that case only one distribution is exists and the other distributions are not important.

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters [9]. The GMM based human detection algorithm is as shown in figure 2.

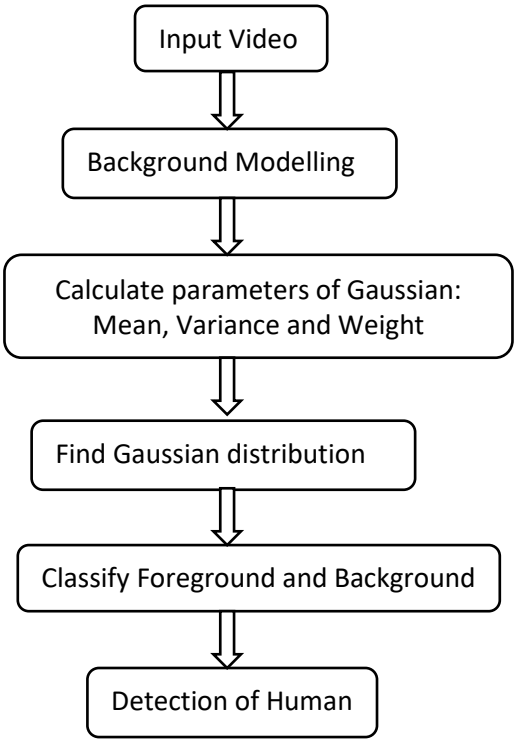


Figure 2: GMM based human detection algorithm

GMM based human detection involves background modelling, Estimation of parameters of K-Gaussian Distribution, Classification of foreground and background.

3.2 Optical Flow Estimation

Optical flow is the pattern of apparent movement of objects, surfaces and edges in a visual scene produced by the relative movement between a spectator and a scene. Lucas- Kanade optical flow method assumes that the flow is basically constant in a local neighborhood of the pixel under consideration, and solves the elementary optical flow equations for all the pixels in that neighborhood by the least squares criterion. By combining information from numerous nearby pixels, the Lucas–Kanade method can resolve the inherent ambiguity of the optical flow equation.

The Lucas–Kanade method makes an assumption that the displacement of the image contents between two consecutive frames is small and almost constant within a neighborhood of the point *p* under consideration. Thus the optical flow equation can be expected to hold for all pixels within a window centered at point *p*.

3.3 SIFT Feature Extraction

There are interesting points on the object that can be extracted to provide a feature description of the object. This description can later be used to locate the object in an image having many other objects. SIFT features provide a set of features of an object that are not affected by many of the complications experienced in other methods such as object scaling and rotation. The SIFT approach takes an image and transforms it into a large collection of local feature vectors [11]. To aid the extraction of these features the SIFT algorithm applies a four stage filtering approach:

1. Scale-Space Extrema Detection

This stage tries to recognize those locations and scales that are recognizable from diverse views of the same object. This can be proficiently accomplished using a "scale space" function. Further, it has been shown under reasonable assumptions it must be based on the Gaussian function. The scale space is defined by the function:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (1)$$

where * is the convolution operator, $G(x, y, \sigma)$ is a variable-scale Gaussian and $I(x, y)$ is the input image.

Various techniques can then be used to detect stable keypoint locations in the scale-space. Difference of Gaussians (DoG) is one such technique, locating scale-space extrema, $D(x, y, \sigma)$ by computing the difference between two images, one with scale k times the other. $D(x, y, \sigma)$ is then given by:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (2)$$

To detect the local maxima and minima of $D(x, y, \sigma)$ each point is compared with its 8 neighbors at the same scale, and its 9 neighbors up and down one scale. If this value is the minimum or maximum of all these points then this point is an extrema.

2. Keypoint Localization

This stage eliminates more points from the list of keypoints by finding those that have low contrast or are poorly localized on an edge. This is achieved by calculating the Laplacian value for each keypoint found in stage 1. The location of extremum **z**, is given by:

$$Z = \frac{\partial^2 D^{-1}}{\partial x^2} \frac{\partial D}{\partial x} \quad (3)$$

If the function value at **z** is below a threshold value then this point is excluded. This removes extrema with low contrast. To eliminate extrema based on poor localization it is noted that in these cases there is a large principle curvature across the edge but a small curvature in the perpendicular direction in the difference of Gaussian function. If this difference is below the ratio of largest to smallest eigenvector, from the 2x2 Hessian matrix at the location and scale of the keypoint, the keypoint is rejected.

3. Orientation Assignment

This step assigns a consistent orientation to the keypoints based on local image properties. The keypoint descriptor can then be represented relative to this orientation, achieving invariance to rotation. The method taken to find an orientation is:

1. Use the keypoints scale to select the Gaussian smoothed image L
2. Compute gradient magnitude, *m*

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \quad (4)$$

3. Compute orientation θ

$$\theta(x,y) = tan^{-1}\big(L(x+1,y) - L(x-1,y)\big) + \big(L(x,y+1) - L(x,y-1)\big) \quad (5)$$

4. Form an orientation histogram from gradient orientations of sample points
5. Locate the highest peak in the histogram. Use this peak and any other local peak within 80% of the height of this peak to create a keypoint with that orientation.

4 Keypoint Descriptor

The local gradient data is also used to create keypoint descriptors. The gradient information is rotated to line up with the orientation of the keypoint and then weighted by a Gaussian with variance of 1.5 * keypoint scale. This data is then used to create a set of histograms over a window centred on the keypoint. Keypoint descriptors typically uses a set of 16 histograms, aligned in a 4x4 grid, each with 8 orientation bins, one for each of the main compass directions and one for each of the mid-points of these directions. This result in a feature vector containing 128 elements. These resulting vectors are known as SIFT keys and are used in a nearest-neighbors approach to identify possible objects in an image.

4.1 Support Vector Machine (SVM) Classifier

A Support Vector Machine (SVM) is a discriminative classifier formally well-defined by a separating hyperplane. In other words, given labeled training data (*supervised learning*), the algorithm outputs an optimal hyperplane which categorizes new examples. In this algorithm, each data item as a point in n-dimensional space (where n is number of features) is plotted with the value of each feature being the value of a particular coordinate. Then, classification is achieved by finding the hyper-plane that differentiates the classes very well.

Two separate hyperplanes on each side are generated as it tries to find that plane which provides the maximum distance between the two parallel hyperplanes [9] [10].

The notation used to define a hyperplane is given by,

$$f(x) = \beta_0 + \beta^T x \quad (6)$$

where β is known as the *weight vector* and β_0 as the *bias*.

The optimal hyperplane can be signified in an infinite number of different ways by scaling of β and β_0. As a matter of convention, among all the possible representations of the hyperplane, the one chosen is

$$\beta_0 + \beta^T x = 1 \quad (7)$$

where x symbolizes the training examples closest to the hyperplane. In general, the training examples that are closest to the hyperplane are called **support vectors**. This representation is known as the **canonical hyperplane** [14]**.**

The distance between point x and hyperplane (β, β_0) is given by,

$$Distance = \frac{|\beta_0 + \beta^T x|}{||\beta||} \quad (8)$$

In particular, for the canonical hyperplane, the numerator is equal to 1 and the distance to the support vectors is

$$Distance_{Support\ Vectors} = \frac{|\beta_0+\beta^T x|}{||\beta||} = \frac{1}{||\beta||} \quad (9)$$

Margin M is twice the distance to the closest examples given by,

$$M = \frac{2}{||\beta||} \quad (10)$$

The problem of maximizing $M$is equivalent to the problem of minimizing a function $L(\beta)$ subject to some constraints. The constraints model the requirement for the hyperplane to classify correctly all the training examples x_i . Formally,

$$min_{\beta,\beta_0} L(\beta) = \frac{1}{2} ||\beta||^2 \text{ subject to } y_i(\beta_0 + \beta^T x_i) \geq 1 \quad (11)$$

where y_i represents each of the labels of the training examples.

5 Experiment, Results and Discussions

The investigation of the proposed approach was implemented on two public video datasets: KTH and Weizmann. The KTH video dataset consists of six human actions: walking, jogging, running, boxing, hand waving and hand clapping. This dataset is also captured over homogenous background and captured using a static camera with 25fps frame rate. The sequences are down sampled to a spatial resolution of 160x120 pixels and have an average length of four seconds. The actions are performed by 25 people for several times in different scenarios. The sample video frames of KTH video dataset with actions is as shown in figure 3.

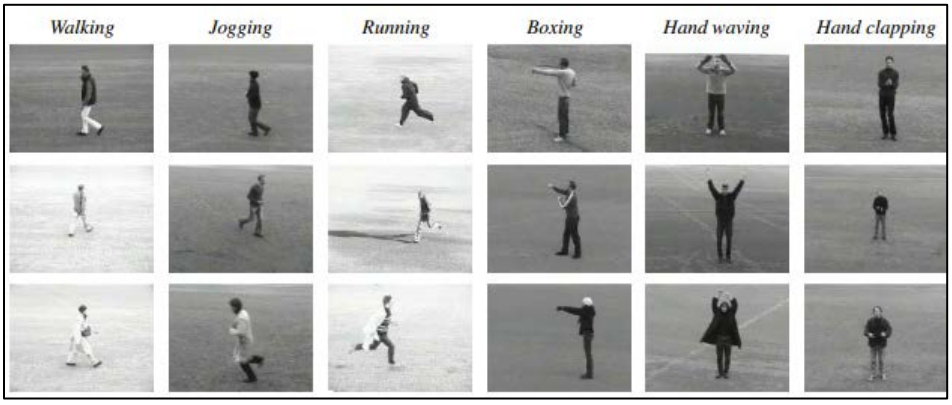


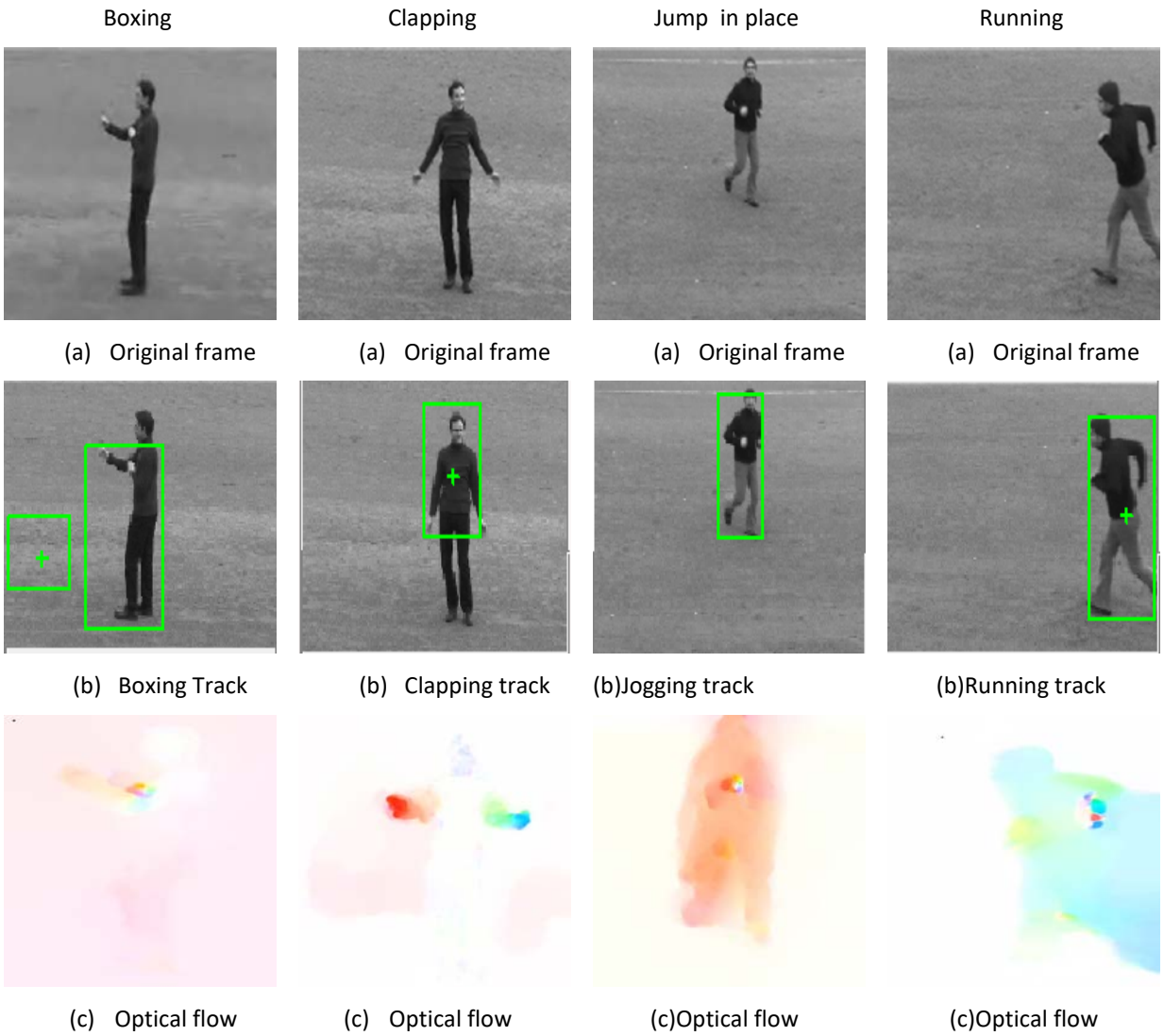
Figure 3: Sample video frames of KTH video dataset

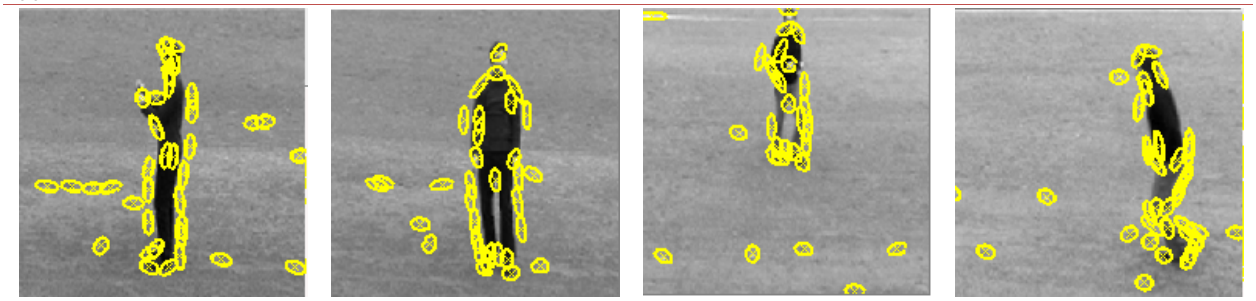
Weizmann dataset consists of 90 video sequences with resolution 180x144 and nine persons have performed ten different actions: bending, jumping, forward jump, jumping in place, running, sideways, skip, walk, one hand wave and two hand wave. There are 90 videos of resolution 180x144 and recorded with a static camera and with homogeneous outdoor backgrounds

The sample video frames of Weizmann video dataset with actions is as shown in figure 4.

Figure 4: Sample video frames of Weizmann video dataset

The complete processing of detection, tracking, optical flow and SIFT features is as shown in figure 5.





(d) SIFT feature (d)SIFT Feature (d) SIFT Feature (d) SIFT feature

Figure 5: Detection, Tracking, Optical flow and SIFT feature extraction for KTH dataset

The classification result of the proposed approach for KTH dataset is portrayed in table 1. The purpose of this depiction is to show the performance efficiency of proposed classification model.

Table 1: Classification accuracy of proposed approach for KTH dataset

	Walk	Jogging	Run	Boxing	Hand clap	Hand wave
Walk	0.924	0.076				
Jogging		0.914	0.086			
Run		0.07	0.930			
Boxing				0.941	0.059	
Hand clap					0.910	0.09
Hand Wave					0.07	0.93

Based on the experimental study, classification accuracy of 92.48% is achieved for KTH dataset. Because of moderate similarity between the features, few actions are misclassified. The classification correctness in terms of percentage is as depicted in the figure 6.

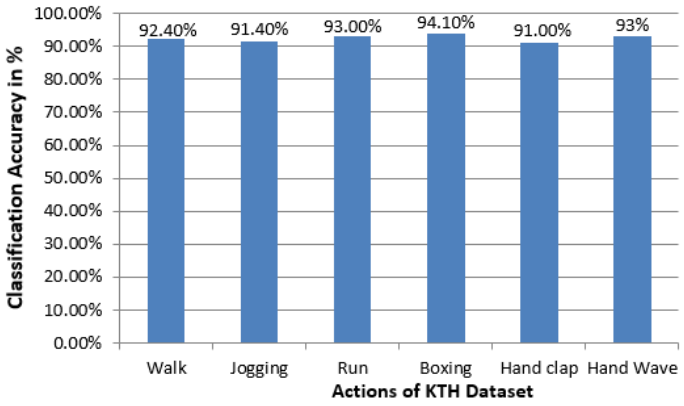


Figure 6: Classification correctness for KTH dataset

The complete processing of detection, tracking, optical flow and SIFT features is as shown in figure 7.

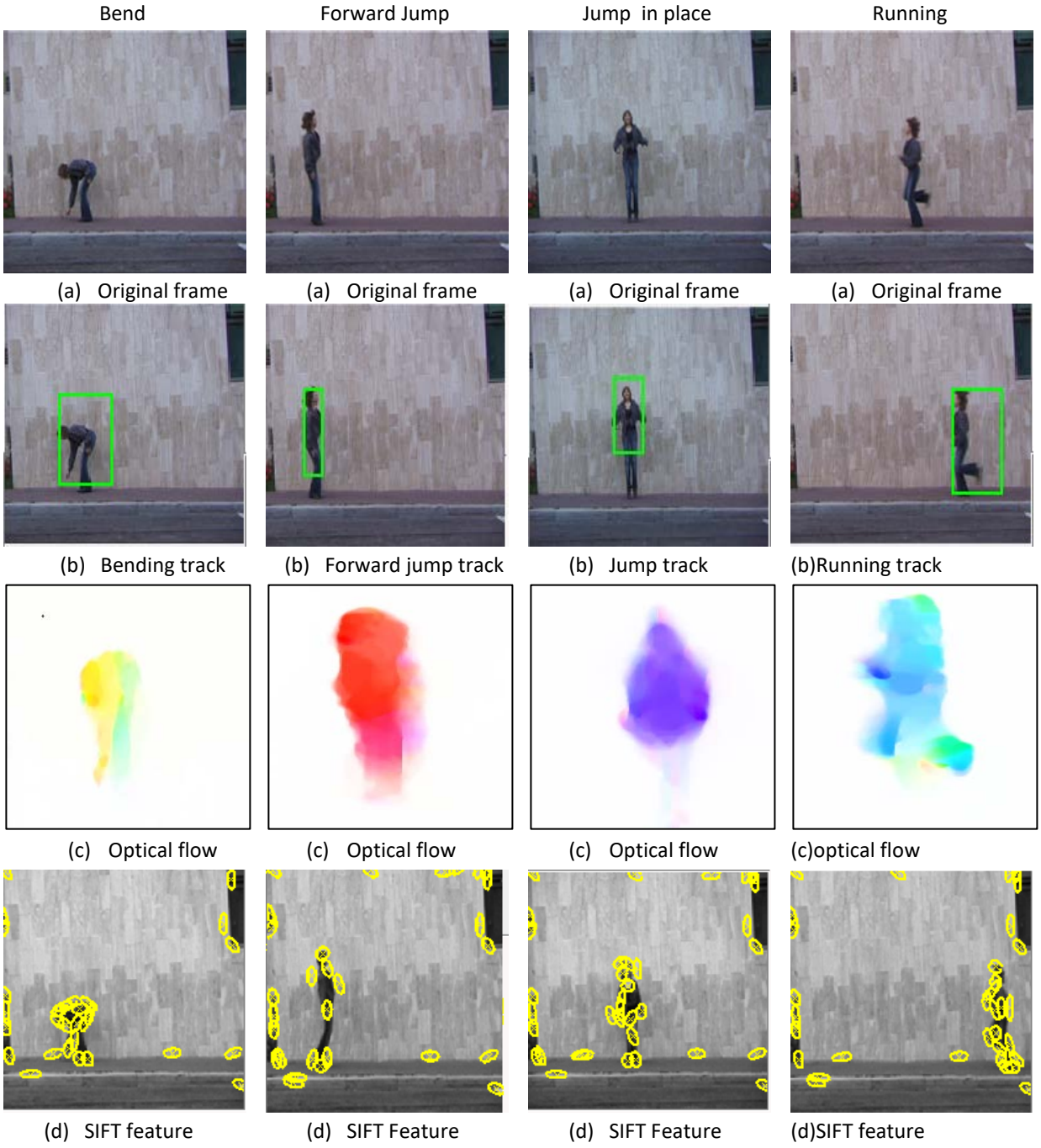


Figure 7: Detection, Tracking, Optical flow and SIFT feature extraction for Weizmann dataset

The classification precision of the proposed approach for Weizmann dataset is presented in table 2. The purpose of this depiction is to show the performance efficiency of proposed classification model.

Table 2: Classification accuracy of proposed approach for Weizmann dataset

	Bend	Jump	Run	PJump	Skip	Side	Jack	Walk	Wave 1	Wave 2
Bend	0.970			0.03						
Jump		0.902			0.098					
Run		0.070	0.930							
PJump				0.935			0.065			
Skip					0.942			0.058		
Side					0.06	0.940				
Jack	0.07						0.930			
Walk			0.05					0.950		
Wave 1									0.922	0.078
Wave 2							0.057			0.943

Based on the investigational study, classification correctness of 93.64% is achieved for Weizmann dataset. Because of moderate resemblance between the features, few actions are misclassified. The classification correctness in terms of percentage is as depicted in the figure 8.

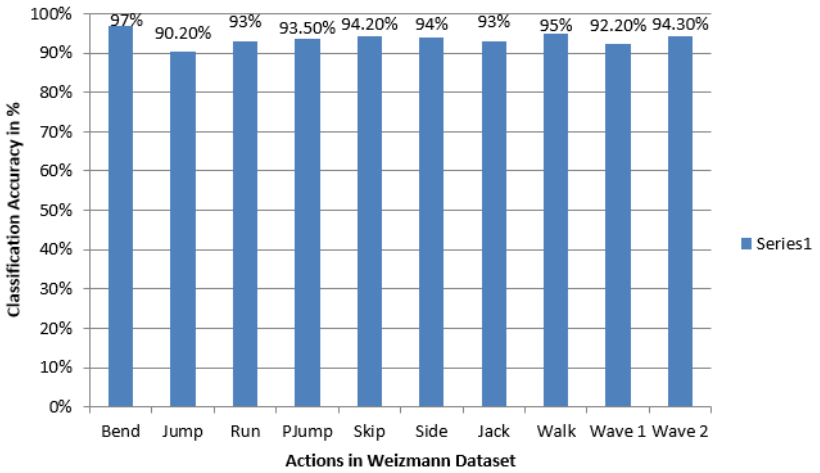


Figure 8: Classification correctness for Weizmann dataset

The same process is repeated for all the other video actions of both KTH and Weizmann dataset and the classification of SVM is done using SVM classifier.

6 Conclusion

This paper presents an activity detection and recognition framework for detection, tracking, feature extraction and classification of actions. The experiment is conducted with two public datasets KTH and Weizmann video datasets. The classification accuracy for KTH dataset is 92.48% and for Weizmann dataset the classification accuracy is 93.64%. The classification precision rate can be improved by using generic and enhanced feature extraction techniques and neural networks for classification. Developing improved optical flow systems apt for large realistic video datasets is imperative to improve the performance of action recognition systems.

REFERENCES

[1] Dadi, H.S., Pillutla, G.K.M. & Makkena, M.L. Ann. Data. Sci. (2018) 5: 157. https://doi.org/10.1007/S40745-017-0123-2

[2] Harihara Santosh Dadi, Gopala Krishna Mohan Pillutla, Madhavi Latha Makkena. "Face Recognition and Human Tracking using GMM, HOG and SVM In Surveillance Videos", Annals of Data Science, 2017

[3] Kale, Kiran & Pawar, Sushant & Dhulekar, Pravin. (2015). Moving Object Tracking Using Optical Flow and Motion Vector Estimation. 1-6. 10.1109/Icrito.2015.7359323.

[4] Kiran Kale, Sushant Pawar, Pravin Dhulekar. "Moving object tracking using optical flow and motion vector estimation", 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 2015

[5] Jin, Ruichen, and Jongweon Kim. "Tracking feature extraction techniques with improved SIFT for video identification", Multimedia Tools and Applications, 2015.

[6] Mona M.Moussa, Elsayed Hamayed ,Magda B.Fayek, Heba A.El Nemr, An Enhanced Method For Human Action Recognition, Journal of Advanced Research, Volume 6, Issue 2, March 2015, Pages 163-169

[7] Dhulavvagol P.M., Kundur N.C. (2018) Human Action Detection and Recognition using SIFT and SVM. In: Nagabhushan T., Aradhya V., Jagadeesh P., Shukla S., M.L. C. (eds) Cognitive Computing And Information Processing. CCIP 2017. Communications in Computer and Information Science, Vol 801. Springer, Singapore

[8] "Cognitive Computing and Information Processing", Springer Nature, 2018

[9] Santosh Kumar, Sanjay Kumar Singh. "Automatic identification of cattle using muzzle point pattern: a hybrid feature extraction and classification paradigm", Multimedia Tools and Applications, 2016

[10] S. Aadhirai, D. Najumnissa Jamal. "Feature extraction and analysis of renal abnormalities using fuzzy clustering segmentation and SIFT method", 2017 Third International Conference on Biosignals, Images and Instrumentation (ICBSII), 2017

[11] Davar Giveki, Mohammad Ali Soltanshahi, Gholam Ali Montazer. "A new image feature descriptor for content based image retrieval using scale invariant feature transform and local derivative pattern", Optik – International Journal for Light and Electron Optics, 2017

[12] C. Schuldt, L. Laptev, and B. Caputo,“Recognizing human actions a local SVM approach” In ICPR, 2004

[13] Junior, Oswaldo Ludwig, et al, "Trainable classifier-fusion schemes: An application to pedestrian detection, "Intelligent Transportation Systems, 2009. ITSC'09. 12th International IEEE Conference on. IEEE, 2009.

[14] Hassaan Ali Qazi, Umar Jahangir, Bilal M Yousuf, Aqib Noor. "Human action recognition using SIFT and HOG method", 2017 International Conference on Information and Communication Technologies (ICICT), 2017

[15] M. N. Al-Berry, Mohammed A.-M. Salem, H. M. Ebeid, A. S. Hussein, Mohamed F. Tolba. "chapter 14 Directional Multi-Scale Stationary Wavelet-Based Representation for Human Action Classification", IGI Global, 2017