International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 3.45 (SJIF-2015), e-ISSN: 2455-2584
Volume 3, Issue 05, May-2017

Human Action Recognition by using Dense Trajectories

W.Jeba[1] ,V.Bhanumathi[2],E.Esakki Madura[3], S.Pradeep[4]

[1]*Department of ECE & Anna university, Coimbatore,*
[2]*Department of ECE & Anna university, Coimbatore,*
[3]*Department of CSE & Anna university, Coimbatore,*
[4]*Department of MCA & Anna university, Coimbatore,*

***Abstract*— Human activity recognition in video is the major problem in the computer vision. The human detection and tracking is significant challenges in the field of computer vision and pattern analysis which is used for video surveillances, gait pathologies recognition, robotics, human computer interaction and sport. When the video dense trajectories can be used to detect the fore ground motion by using the optical dense flow method. The dense representation is capturing the local information in the each video. The human motion can be tracking by using the spares Kanade Lucas Tomasi based tracker method.**

The motion and structure descriptors are used to describe the human motions based on the motion boundary descriptors such as histogram of gradient and histogram of optical flow. These descriptors are detecting the human by computing derivates in the horizontal and vertical components of the optical flow. We evaluate the pseudo likelihood estimation of each trajectory that is shape factor and scale factor values in the each video dataset. The human activity recognition is classifying the human action by using K Nearest Neighbor classifier and binary Support Vector Machine classifier. The MATLAB tool is used to implement the human action recognition. Various types of human actions are tested and the performance been analysed. In this project, experimentally three datasets are used such as Human 3.6Million dataset, KTH dataset and Berkeley MHAD dataset.

***Keywords*— Dense Trajectories, Sparse KLT, MBH, HOG, Human Action**

1. INTRODUCTION

The human activity recognition is the process of recognizing the conjurations of the body pose from a single, typically monocular, image. In the computer vision and machine learning techniques, one of the main problems are the human pose estimation that has been studied for well over 15 years. The example for the human pose estimation is the human computer interaction, video Surveillance and activity recognition marker-less motion capture. The Motion Capture (Mo-Cap) technology is used for applications to clinical analysis from character animation of gait pathologies.

The most significant challenges of the human pose estimation is a very difficult and still challenging problems are large variability of human visual appearance in an image such as people's clothes, motions, occlusions, outliers, variability in lightening conditions, variability in human physique, human skeletal structure complexity, high dimensionality of the pose and the 2D planar image cannot cope with the loss of 3D information due to camera view changes.

The human action detection and tracking in video processing has received attention in the last few years. A monocular image of 2D human body pose detecting and tracking human body configurations in unconstrained video is also the challenging problem. This is the impractical and time-consuming. Since, the vision-based human motion analysis to understand the movements of the human body.

The movements of the human body can be interpreted on a physical level that is pose estimation and the movements of the human body over with the time that is action recognition. The behaviour recognition is a challenging problem in the computer vision, where used in the fields of Medical Diagnosis, Military systems, Document analysis, Manufacturing, entertainment and sport and active assisted living.

The human activity recognition is represented the different types of categories in computer vision and pattern analysis, such as Silhouette based Representation, Body Parts /Joints based Representation, Space-Time Interest Point's Representation, Depth Map based Representation. These are representations mentioned below the human action recognition and also behaviour recognition including human pose estimation.

The silhouette or contour or shape based representation is an effective representation of the shape of the human body postures to be described. The human pose estimation is drawn lines on the silhouette representation. This representation can be divided into two ways of the human behaviour recognition. One is to be extracting the behaviour descriptors from the frame sequence of silhouettes. Another one is to be extracting the features from each silhouette. The human joints or body parts based representation usually acquired data by using the markers to the subjects and to get the 3D position. These human body parts are representing in the human action recognition under realistic imaging conditions. A single person specific body part model which includes different 3D body joints such as head, torso, neck, left shoulder, right shoulder, left elbow, right elbow, left hand, right hand, left hip, right hip, left knee, right knee, left feet and right feet. These are invariant to the body size, orientation and human position.

The STIP represent in to extend the notation of the spatial temporal domain and reflecting the interested point that can be used to represent the video data. We estimate both space and time information to detect and track the human object recognition. The depth map motion capture image features from the space-time depth difference image are obtained from the hierarchical of the silhouette box.

2. RELATED DATAEST

In the paper work has experienced with different dataset in the human activity recognition in video processing over the state-of-the-art approaches such as KTH Dataset, Human 3.6M (Million) Dataset, Berkeley MHAD (Multi-Model Human Action Dataset). These are dataset videos are used in the project that is detecting and tracking human body configurations in the unconstrained video is still challenging problem.

2.1 KTH Dataset

The human activity video dataset contains number of action such as walking, jogging, running, boxing, hand waving and hand clapping. The homogeneous indoor/ outdoor back grounds performed by 25 persons, 4 scene and 6 verbs with multi class recognition accuracy.

2.2 Human 3.6M Dataset

The human 3.6Million contains 3D human poses and corresponding images, diversity and size, accurate capture and synchronization includes high resolution, accurate 3D joint position and joint angles from high speed motion capture system, pixel level 24 body part labels for each configuration, time of flight range data, 3D laser scans of the actors and accurate background subtraction, person bounding box. The support for development includes with pre computed image descriptors, software for visualization and discriminative human pose prediction and performance evaluation on withheld test set. This Human 3.6M dataset have capturing the videos by using the 4 cameras and to detect the human activity.

2.3 Berkeley MHAD Dataset

The comprehensive Berkeley Multi-Model Human Action Database contain 11 actions performed by 7 male and 5 female subjects in the range 23-30 years of age expect for one elderly subject. All the subjects are performed 5 repetitions of each action, yielding about 660 action sequences which correspond to about 82 minutes if total recording time. These dataset videos captured by two different cameras. The various human actions are jumping, clapping, sitting, standing, sit stand, jumping jacks, walking, running, one wave arm, two wave arm, punching and throwing, etc.

3. RELATED WORKS

Feng Zhou (2016) proposed to match the motion of 3D model for estimating the camera view and for selecting the subset of video trajectories [1]. The approach is to solve the correspondence between the video sequence and 3D motion capture data by using the STM (Spatio-Temporal Matching). The trajectory based representation of the paper was explained the Sparser KLT (Kanade Lucas Tomasi) feature tracker, because of the determined the foreground motion accurately in which video fast irregular motion. Spatio-temporal bilinear basis explained here includes learning the DCT (Discrete Cosine Transform) bases independently foe each frame and Linear bases independently foe each video frame. The STM learns a bilinear spatio-temporal 3D model from motion capture data that will be used to constraints possible video trajectories. The disadvantage is the highly computational cost which is computed independently for each frame and for calculating the joint's response. Zhe Jio proposed to develop the 3D MoSIFT (Motion Scale Invariant Feature Transform) for the depth description and motion information [2]. The main advantage is HMM used to improve the accuracy of human behaviour recognition and Individual movements are very efficiently. The disadvantage is Combination of MoSIFT & HMM can make too much noises being introduced. Heng Wang proposed to represent the good coverage of foreground motion and local motion information for action recognition of the video [4]. The structure and motion boundary descriptors are computed in a 3D volume interest points into its horizontal and vertical components of the optical flow. The disadvantage is the video description not limited to Bag-of-features representation and performance is currently limited by the optical flow descriptors.

Angela Yao proposed to combined the human action detection and a deformable part model to estimate the 3D poses from the multiple activity recognition [4]. The advantage is Low dimensional embedding is efficient. Minimum no. of particles is used. The disadvantage is Human-object interaction is not efficiently worked. Ross Messing proposed to shows the velocity history feature can be extended both latent velocity model and high level semantic model [5]. The advantage is Capturing the non-local structure is more valuable than the spatio - temporal method. The disadvantage is activity recognition will face the more complicated activities; The Greater computational cost and larger resolution video is very complex activities.

Vennila Megavannan proposed to recognize human action by using the depth map images from kinect camera [7].The main advantage is reducing the complexity of the silhouette extraction and increase the robustness of the action recognition and representation. The disadvantage is more depth variation of the silhouette regions cannot be extracted. Lubomir Bourdev proposed to detect and segment the human pose estimation of an image. There are two criteria i.e., detect the poselet of a given images, another one is localize the 3D joint configuration of the human [8]. The main disadvantage is does not have the quantitative data that is soft region labelling in the poselet.

Bruce Xiaohan Nie proposed to integrate training and testing of the two models such that one is spatio temporal (ST) parts model and graph model [9]. The limitation is does not handle the self occlusion explicitly for human pose estimation. Jasreet Kaur proposed to detect the human action by using region of interest (ROI) part and morphological operator. The recognizing the human movement with the help of angle value, correlation and interpolation.[10]. The advantage is very high accuracy of recognizing the human activity within the videos. The disadvantage is does not processed then training images are given with different classifiers of the human action recognition.

4 OPTICAL FLOW METHOD

The optical flow method is defined as the pattern of apparent motion of the human, objects, edges in the visual scene caused by the motion between the camera view and video sequence. On the other hand, the optical flow method is the distribution of the apparent velocity of the object in the each video trajectories measured.

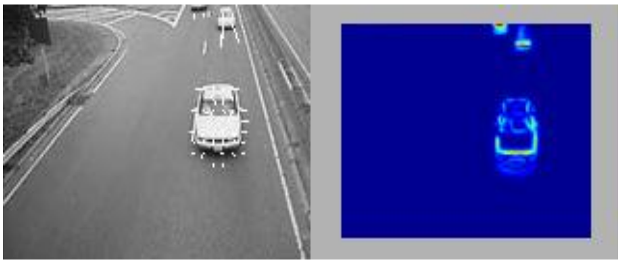


Figure 1 Optical Flow Estimation

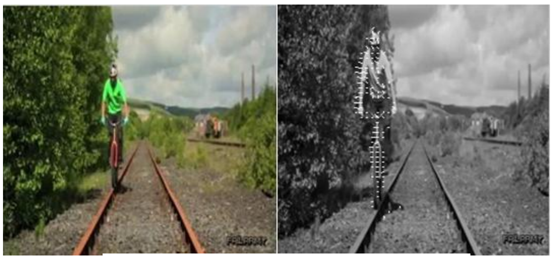


Figure 2 Sparse Optical Flow Estimation

For the motion estimation in the video stream or motion based human detection and tracking system as shown in Figure 1, the optical flow method has a two following method, such that Sparse Optical Flow Method, Dense Optical Flow Method. These two methods are used in the computer vision application; optical flow estimation is moving object will be more apparent than the distant object at the same speed. This technique used to estimation the motion vectors in the each frames of the video sequence.

4.1 Sparse Optical Flow Method

One of the Sparse Optical Flow Method is the Lucas-Kanade method. This method used to process only the some interesting features such as points, edges, corners, etc., from the each frame of the video sequence representing in the Figure 2. Commonly Lucas-Kanade method is very accurate and generally faster for real-time application.

4.2 Dense Optical Flow Method

One of the Dense Optical Flow Method is the Gunner Farneback's Optical Flow Method. This method used to process all the pixels. The advantage of dense optical flow method can be estimated the more pixels than the sparse optical flow method as shown in Figure 3. The Dense optical flow method is more accurate, but slower than the sparse method.

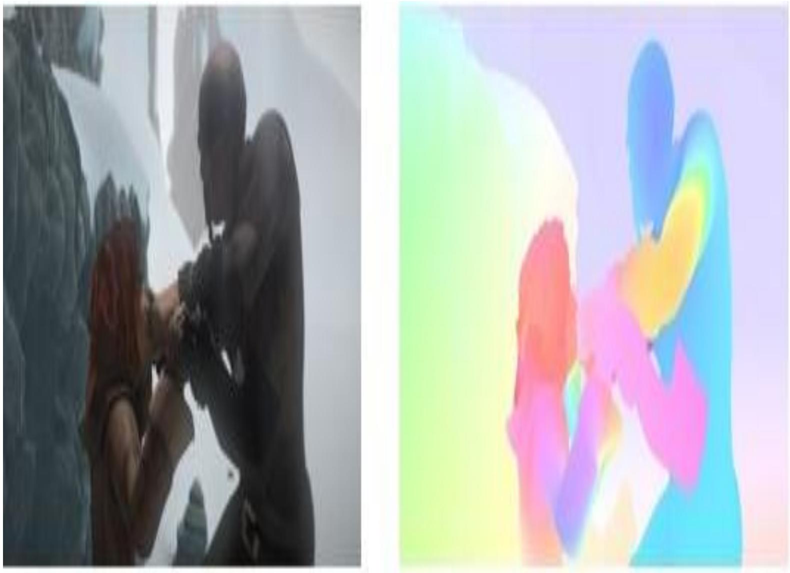
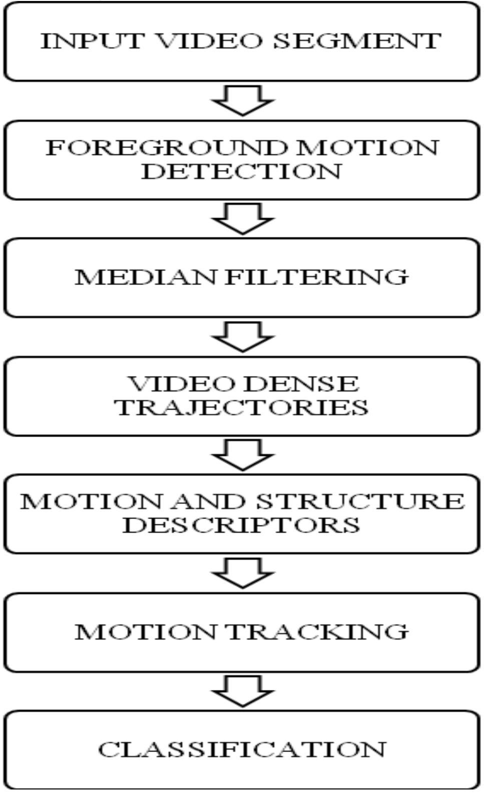


Figure 3 Dense Optical Flow Estimation

5. SYSTEM OVERVIEW

The overview of this approach is illustrated in the block diagram. In order to represent the human activity recognition in video processing has the various blocks in the system is as follows.



Block Diagram of Human Activity Recognition in Video

The block diagram of the human activity recognition in the video containing the fore ground motion detection, median filtering, video dense trajectories, motion and structure descriptors, motion tracking and K-nearest neighbor classification and also the binary support vector machine classification. The each block can be explained in the following sections in the human detecting the deformable parts and human tracking the outliers, occlusions, missing data and mismatches of the video.

5.1 FORE-GROUND MOTION DETECTION

The foreground motion detection or background subtraction is defined as the techniques in the fields of image processing and computer vision, the frame's foreground is extracted for the further processing such as object recognition, etc. Generally, the frame could be assigned the region of interest objects, the background subtraction for detecting the moving objects in video segment. Many applications could not need to know everything about all the moving video sequence, but we required only the changing information in the video sequence.

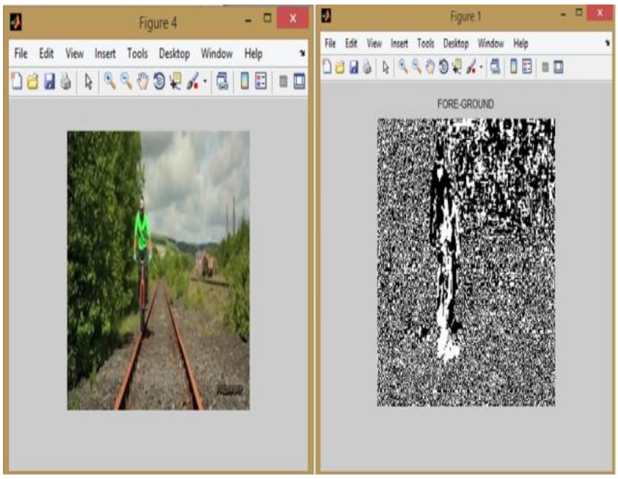


Figure 4
File Edit View Insert Tools Desktop Window Help
Figure 1
File Edit View Insert Tools Desktop Window Help
FORE-GROUND

Figure 4 Fore ground Motion Detection

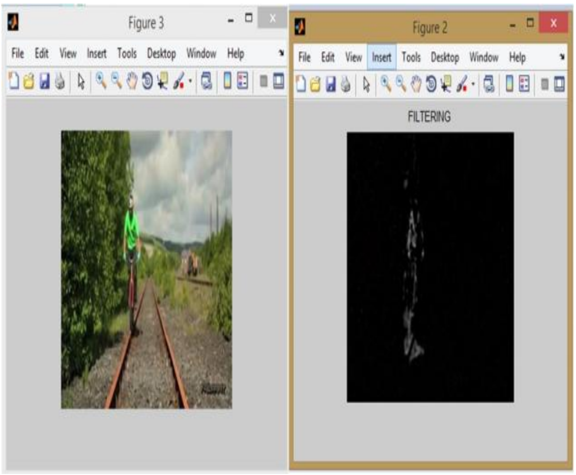


Figure 3
File Edit View Insert Tools Desktop Window Help
Figure 2
File Edit View Insert Tools Desktop Window Help
FILTERING

Figure 5 Median Filtering of video frame

The figure 4 represents the foreground motion detection in the video dataset. The different video frames are used and only foreground motion is extracted from the successive frames.

5.2 MEDIAN FILTERING

The median filtering allows maintain the sharp motion boundary. It is more robust to outliers than bilinear interpolation and improving the video trajectories at the motion boundary. Compared to the bilinear interpolation, the motion boundaries are blurred and foreground motion and back ground motion are confused. The median filtering is a nonlinear digital filtering techniques used to removing the noise. The median filtering of the video frame are as shown in figure 5. This is the image pre-processing step in the edge detection in the foreground object detection.

5.3 VIDEO DENSE TRAJECTORIES

The dense optical flow field is computed and the trajectories can be tracked very densely. It is the smoothness which allows more robustness tracking of the irregular motion video. When the two frames are sampled at different time and finding the change position of the each pixel in the first and next frame in the video sequence. The polynomial expansion is to approximate the pixel velocity in the neighborhood successive frames throughout the videos. The video dense trajectories are major feature extracted the foreground motion detection exactly. For each video segment, we extract the video trajectories densely sampling features points in the first frame and track the next frame by using dense optical flow method. In this section, we study how to extract the dense video trajectories and evaluate the pseudo likelihood value of the each video trajectory belong to the 3D joint independently of the each video frame.

5.3.1 Dense Sampling Feature Points

First sampling is carried out on the each spatial scale separately. The features points are equally cover over all spatial position and scales. The spatial scale factor is increasing the factor of 1/V2. Our main goal is to track all the sampled feature points throughout the video frames as shown in Figure 6

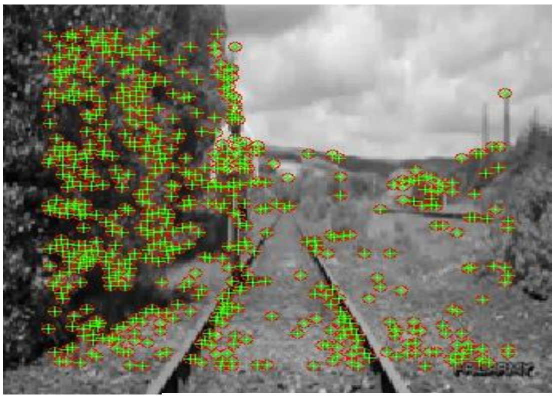


Figure 6 Densely Sampling Feature Points

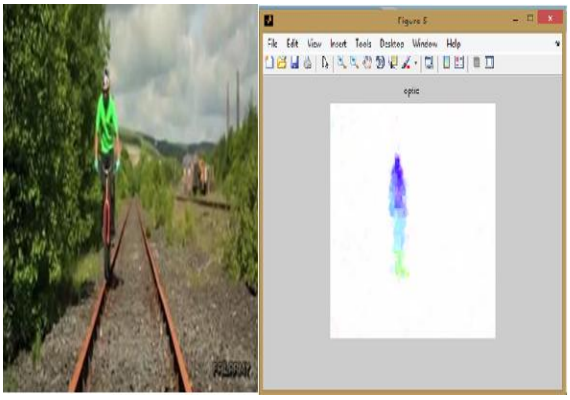


Figure 7 Video Dense Trajectories

Then, we set the threshold value of the each n frame in the video segment such that

$$T=0.001*\max[\min(\lambda_n^1,*\lambda_n^2)] \qquad (3.1)$$

Where, $(\lambda_n^1, \lambda_n^2)$ are the Eigen values of the features points in the n^{th} frame and 0.001 indicate the good compromise between density and saliency of the densely sampled feature points. When the homogenous area can been removed.

5.3.2 Dense Trajectories

The dense trajectories in the example representation of Figure 7 for each video frame I_t , its compute the optical flow field w = (u, v) with respect to the next frame of the I_{t+1} ,that is

$$I_x u + I_v v + I_t = 0 \qquad (1)$$

Where, u is the horizontal optical flow components,
v is the vertical optical flow components and
I_x, I_y, I_t are the spatio temporal Image (frame) brightness derivates.

Given a point $P_t = (x_t , y_t)$ in frame It, its tracked position in frame It+1 is smoothed by using the median filter on w:

$$P_{t+1} = (x_{t+1} , y_{t+1}) = (x_t , y_t) + (M * w)\, l_{(xt\, ,\, yt)} \qquad (2)$$

Where, M is the median filtering kernel.

The median filter kernel M size is the 3*3 pixels. The median filter is used to track the outliers than the bilinear interpolation motion model between the two successive frames. It assumes that the optical flow is essentially constant velocity in the neighbourhood pixel value in each video frames and solve the basic optical flow equation by using the least square criterion.

5.4 MOTION & STRUCTURE DESCRIPTORS

The motion descriptor is a spatio temporal representation of motion estimation in the optical flow and moving camera views. The local motion descriptors are HOG (Histogram of Gradient), HOF (Histogram of Optical Flow) descriptors as shown in Figure 3.8. These descriptors computed with the dense trajectories.

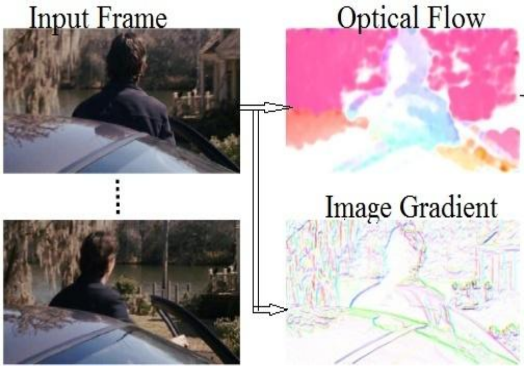


Figure 8 Motion and Structure Descriptors

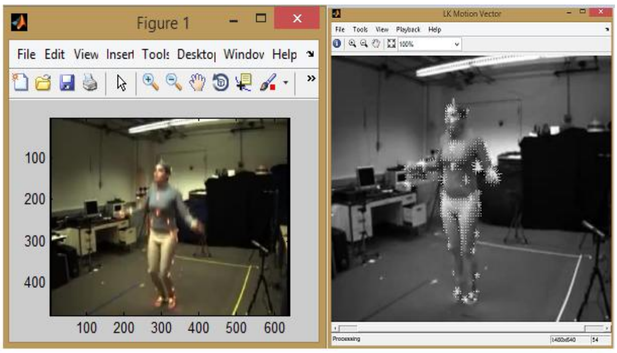


Figure 1
File Edit View Insert Tools Desktop Window Help
100
200
300
400
100 200 300 400 500 600
LK Motion Vector
File Tools View Playback Help
100%
Processing

Figure 3.9 MHAD dataset for Jumping Motion Tracking

The motion boundary descriptors are detecting the human by computing derivates in the horizontal and vertical components of the optical flow. The optical flow shows the constant motion in the back ground and the motion boundary descriptors shows the relative motion between the foreground and back ground. The image gradients or flow orientations are indicated by the color that is hue and magnitude that is saturation. The gradient of the optical flow represents the Motion Boundary Histogram (MBH) descriptors. Thus, the result can be computed with motion boundary descriptors and gradient optical flow in the video segment.

5.5 MOTION TRACKING

The human or object can be tracked by using the Lucas Kanade based tracker and Sparse Kanade Lucas Tomasi based feature tracker. The motion tracking method used to track the human object inducing the region of interest in the object frame in the video segment. For example representation of MHAD dataset for jumping motion tracking as shown in Figure 9. The motion video vector compared to the sparse KLT based feature tracker that is guarantees of the densely tracking feature points gives the good coverage area of the foreground motion in the faster irregular videos.

5.5.1 Sparse KLT Based Tracker

The Sparse KLT (Kanade-Lucas-Tomasi) feature tracker is an approach to extract the features from the given input videos. The disadvantage of the Lucas Kanade method is less sensitive to the image noise than the point wise method and also purely local method that cannot provide the flow information in the interior of uniform regions of the image or frame. The KLT based tracker makes use of spatial intensity information to direct search for the position, location and orientation that yields the best matches. This technique is faster than the traditional techniques. The point tracker is used for short term tracking as the part of the large term tracking.

5.5.2 Advantage of the Sparse KLT

The advantage of the Sparse KLT based tracker compared to the Lucas Kanade based tracker such that

- Tracking object does not change the shape,
- Good coverage area of the foreground motion,
- Improve the quality of the each video trajectory in the fast irregular motion. The densely tracking is the feature point guarantees of the Sparse Kanade Lucas Tomasi based tracker.

5.5.3 Comparison of the Sparse KLT and Lucas Kanade Method

The comparison results for the Sparse KLT based tracker and Lucas Kanade based tracker based on the cycling action is given below the Figure 3.10 shown.

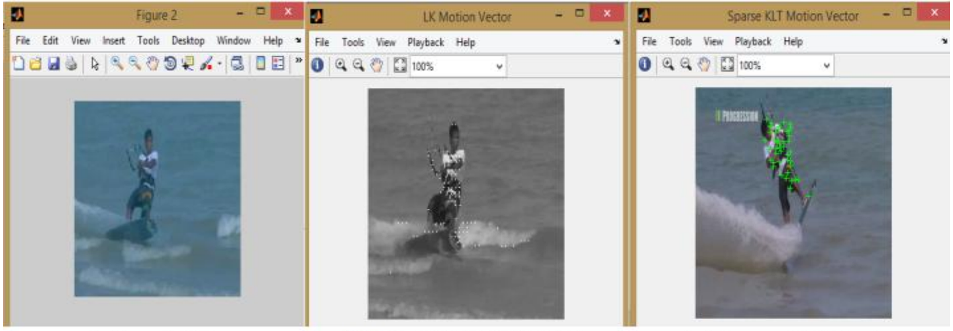


Figure 2
File Edit View Insert Tools Desktop Window Help
LK Motion Vector
File Tools View Playback Help
100%
Sparse KLT Motion Vector
File Tools View Playback Help
100%
PROCESSING

(a) Input Video (b) Lucas Kanade Tracker (c) Sparse KLT Tracker

Figure 10 Comparison Outputs

The given input video can be used to tracking the human or object by using the two different methods such that Lucas Kanade based tracker method and Sparse Kanade Lucas Tomasi based tracker method. The comparison figure 10 shown as the good coverage area of the foreground motion in the fast irregular motion videos.

Accordingly the tracking the object shape cannot be changed and for these exhibit visual texture and the improving the quality of the each video trajectories in the video frame. It is operated as the equally divided the original frame into the smaller section and assumes the constant velocity vector in the each smaller section of the overall video. When the optical flow smooth over the entire frame in the Horn Scheck Method computed based on the two frames sampled at different slight and different time. As find the change the position of each pixel in the first frame and track the next frame in the video segment.

5.6 PSEUDO – LIKELIHOOD EVALUATION

The pseudo likelihood estimation is defined as an approximation to the joint probability distribution of the collection of variables in the 3D joint applied to state of the art approaches. This pseudo likelihood approach was the spatial dependence of the matrix independently on the each frame.

Table 1 Result for feature value in different video dataset

S.N	Video Dataset	Optical flow Elapsed Time	Pseudo Likelihood Value	
			Shape Factor	Scale Factor
1.	Bending	32.101346 sec	87.5709	87.5002
2.	Clapping	33.762715 sec	87.5862	87.5002
3.	Jumping Jacks	31.254153 sec	87.6741	87.5002
4.	Jumping	40.224755 sec	87.7083	87.5002
5.	One arm wave	54.388193 sec	87.8065	87.5002
6.	Two arm wave	40.479487 sec	87.8642	87.5002
7.	Sit stand	26.254873 sec	87.9518	87.5003
8.	Throwing	32.453600 sec	88.0223	87.5002
9.	Punching	25.418602 sec	88.1505	87.5002
10.	Cycling	25.248885 sec	87.6472	87.5002
11.	Surfing	30.738146 sec	88.1965	87.5002
12.	Running	29.004748 sec	86.8679	87.4897
13.	Direction	30.747194 sec	87.7457	87.4985
14.	Sitting	30.785014 sec	87.8451	87.4988
15.	Walking	40.483234 sec	87.5571	87.5002

Given a set of random variables P = P_1, P_2, …,P_{mp} and a set of dependencies between these random variables in the n^{th} frame of the trajectory(mp).

$$\mathbf{P = [\ P_1^1\ \ P_2^1\ \ ....\ \ P_{mp}^1 \\ P_1^2\ \ P_2^2\ \ ....\ \ P_{mp}^2 \\ .\ \ \ .\ \ \ .... \\ P_1^n P_2^n....P_{mp}^n]} \qquad (3)$$

Where, P denotes the 2D coordinate of the mp trajectory in the n^{th} frame. This pseudo likelihood estimation is used for finding the missing data computed more efficiently than the likelihood estimation.

5.7 CLASSIFICATION

The human action recognitions are classified at different classification technique, such as SVM (Support Vector Machine) classification and KNN (k Nearest Neighbor) classification. We have trained for every human action separately in the SVM classification. The KNN classifier is compared to SVM classifier, to detect the human activity in the trained video sequence.

In machine learning, the supervised learning models are support vector machines with associated learning algorithms that recognizing the patterns and analysing the data, used for classification and regression analysis. The k-Nearest Neighbour techniques has been used in various areas such as Computer Vision, Bio- informatics, Multimedia database, document retrievals, Marketing data analysis and image processing and data compression. The construction of the kNN classifier is classified using the ClassificationKNN.fit Matlab command for the given database.

6. RESULTS AND DISCUSSION

The recognition of the human activity or pose estimation has been successful implemented for all types of human actions using MATLAB tool. In this project 15 videos were used for training and 25 videos were used for testing from collect the different dataset such as KTH dataset, Human 3.6 Million dataset, Berkeley MHAD dataset. The various results are shown below.

6.1 RESULTS FOR VARIOUS DATASETS

The following tables shows the results of detecting the foreground motion of the deformable parts by using the video dense trajectories and tracking the outliers, occlusions, missing data and mismatches of the human motion used as Lucas Kanade based tracker method and sparse KLT based feature tracker method. The Motion Boundary Descriptors are the optical flow and image gradient of the histogram shows the following tables. The pseudo likelihood values can be estimated from each video trajectory in the video frame belongs to the 2D joints. And then classification results for the human activity recognition or human pose estimation was analyses. The table 2 represents the feature value in different video dataset and figure 11 represents the feature value in different video frame

Table 2- Result for feature value in different video frame

S.NO	VIDEO FRAME	MEAN	SD	THRESHOLD	TOTAL AREA
1.	BENDING	129.5431	70.8880	0.2000	6686
2.	CLAPPING	130.4986	69.5440	0.2078	35241
3.	JUMPING JACKS	143.1330	55.4741	O.2138	39787
4.	JUMPING	127.4783	73.8187	0.2667	8240
5.	ONE ARM WAVE	127.5266	73.7736	0.2235	29730
6.	TWO ARM WAVE	131.4122	68.4668	0.1922	32648
7.	SIT STAND	127.5367	73.8396	0.2471	9369
8.	THROWING	128.2549	72.7460	0.2157	34426
9.	PUNCHING	128.6920	72.1032	0.2118	34103
10.	CYCLING	135.1062	63.9433	0.3059	36531
11.	SURFING	135.1252	63.8813	0.2196	20558
12.	RUNNING	127.5000	73.8900	0.3373	6612
13.	DIRECTION	137.1890	61.6642	0.3961	61690
14.	SITTING	128.4688	72.3744	0.4431	56282
15.	WALKING	135.1179	63.8759	0.2902	28056

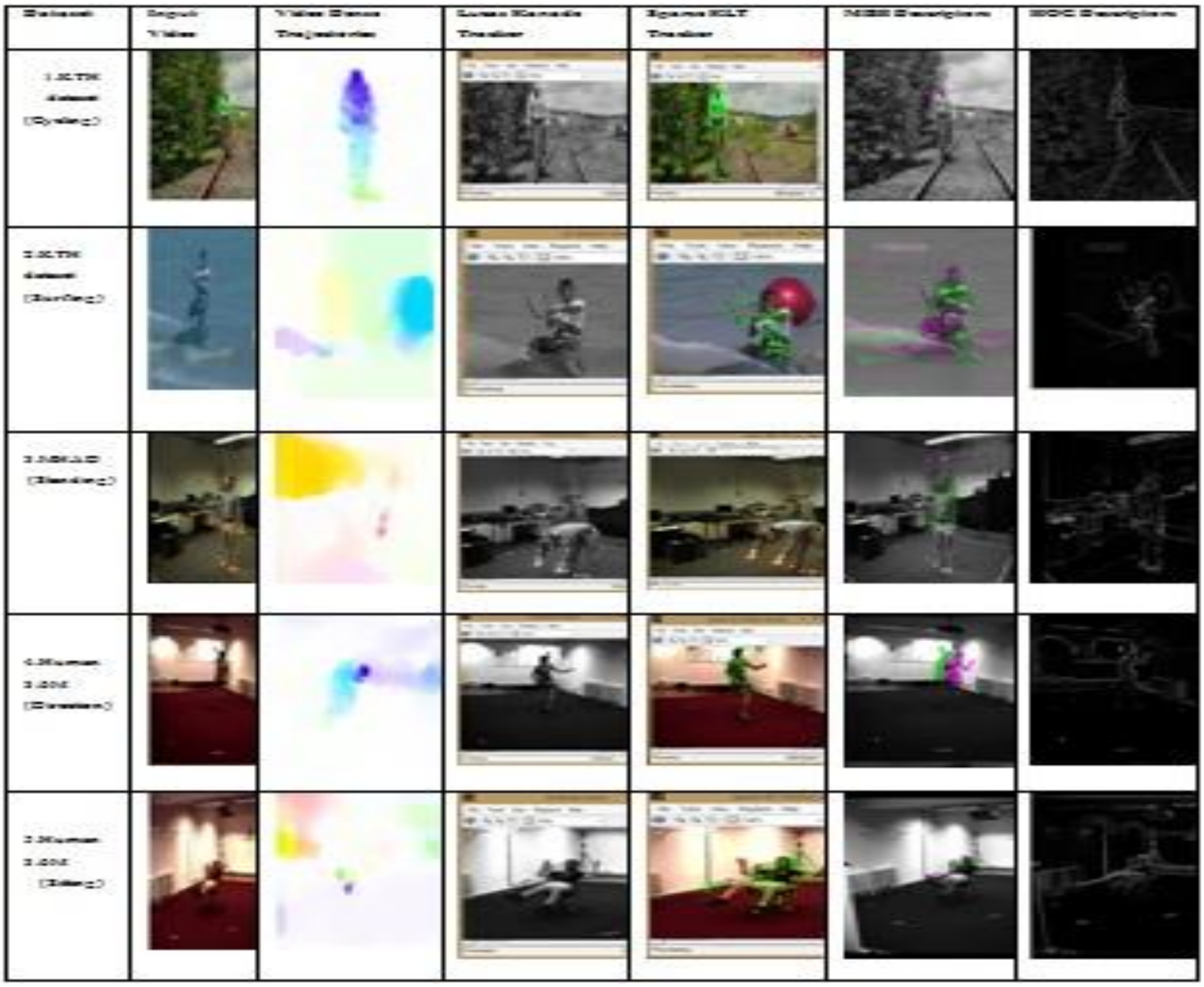


Figure 11 Results for each Video Dataset

Based on the above features, the human actions are classified as using SVM classifier of the each video frame in the different video dataset.

6.2 CLASSIFICATION RESULTS FOR VARIOUS HUMAN ACTIONS

The human activity recognition was classified in the various human action shown in Figure 4.1 such as walking, running, bending, clapping, cycling, jumping, sitting, throwing, one wave arm**,** two wave arm, surfing and punching, etc .

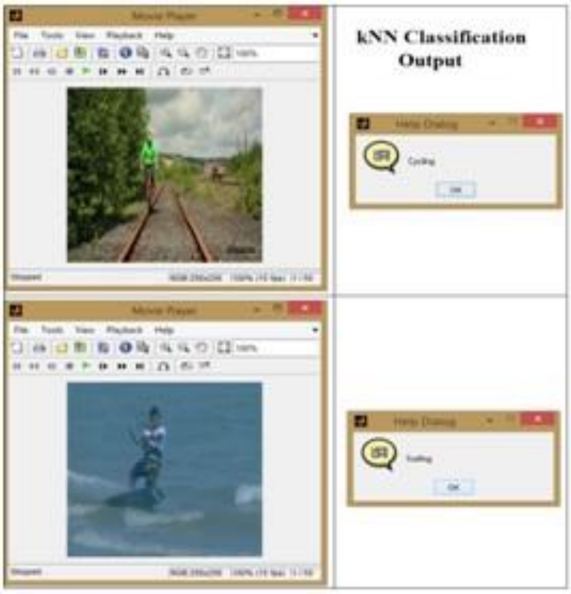
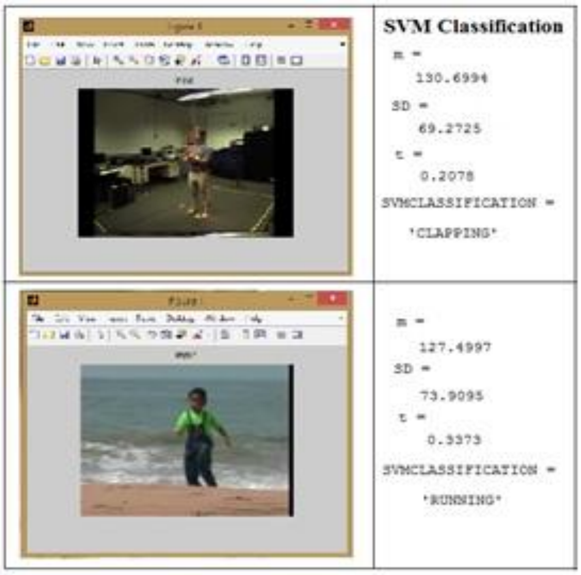


Figure 12 kNN Classification Output



```
m =
   130.6994
SD =
   69.2725
t =
   0.2078
SVMCLASSIFICATION =
   'CLAPPING'
```

```
m =
   127.4997
SD =
   73.9095
t =
   0.3373
SVMCLASSIFICATION =
   'RUNNING'
```

Figure 13 SVM Classification Output

The SVM classification is used to detect the action and classified. The feature values can be determined such as mean, standard deviation and threshold value of the each video frame in the video dataset. The working of the various blocks in the automatic human action classification based on human activity or pose estimation has been explained in detail in this chapter. The experimental analysis and results are done in this chapter. From the results it is proved that the performance of the human activity recognition in different types of human action video dataset is better compared to other existing methods.

7. CONCLUSION

Human detection and tracking in the videos and classification for the different types of human action has been implemented in this work. This method is very effective, since it uses the motion structure descriptors and histogram of the gradient which provides better response than the existing techniques. The K-Nearest Neighbour (kNN) classifications shows output of the dialog box of the human action in the given input videos segments in the computer vision. The future work of this project is to solve the correspondence between the video trajectories and 3D motion capture model for human pose detection and to use multiclass classifiers such as convolutional neural networks (CNN) to classify the human action.

REFERENCES

[1] Feng Zhou and Fernando De la Torre “Spatio-Temporal Matching for Human Pose Estimation in Video” *IE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, August 2016.

[2] Yuexin Wu, Zhe Jio, Yue Ming, Juanjuan Sun, Liujuan Cao “Human Behavior Recognition based on 3D Features a Hidden Markov Models “ *Springer-Verlag* London March 2015

[3] Heng Wang, Alexander Kl□aser, Cordelia Schmid, Cheng-Lin Liu “Dense Trajectories and Motion Bounda Descriptors for Action Recognition “ *National Laboratory of Pattern Recognition, CASIA,* no 8050, August 2012

[4] Angela Yao, Juergen Gall, Luc Van Gool “Coupled Action Recognition and Pose Estimation from multiple view *International Journal on Computer Vision,* vol. 100, no. 1, pp. 16–37, 2012

[5] Ross Messing, Chris Pal, Henry Kautz “Activity Recognition using the Velocity Histories of Tracked Key points” *Pr IEEE 12th International Conference on Computer Vision,* 2009, pp. 104-111

[6] Zhao, Y., Liu, Z., Yang, L., Cheng, H, “Combing RGB and depth map features for human activity recognition”, Sig and Information Processing Association Annual Summit Conference (APSIPA ASC), pp. 1–4 (2012)

[7] Vennila Megavannan, Bhuvnesh Agarwal, R. Venkatesh Babu "Human Action Recognition using Depth Maps " Pr *IEEE Conference on Computer Vision Pattern Recognition,* 2014, pp. 1653–1660

[8] Lubomir Bourdev and Jitendra Malik "Poselet: Body Part Detectors Trained using 3D Human Pose Annotations*" IE Trans. Pattern Analysis and Machine Intelligent.*, vol. 28, no. 1, pp. 44–58, Jan. 2012

[9] Bruce Xiaohan Nie, Caiming Xiong and Song-Chun Zhu "Joint Action Recognition and Pose Estimation from Vide *CVPR* 2015

[10] Jasreet Kaur and Manpreet Kaur, "Human Action Recognition using SVM and KNN Classifiers" Conf. Proc 2 *International Conference on "Recent Innovation in Science, Technology and Management" (ICRISTM-16)* August 201

[11] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively train part-based models," *IEEE Transaction on Pattern Analysis and Machine Intelligence.*, vol. 32, no.9, pp. 1627–1645, S 2010

[12] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transaction on Patte Analysis and Machine Intelligence.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013

[13] Leonid Sigal, et al., "Human Pose Estimation" *IEEE Conference on Computer Vision and Pattern Recognition*, 2010

[14] H. Wang, A. Kl€aser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for acti recognition," *International Conference on Computer Vision,* vol. 103, no. 1, pp. 60–79, 2013.

[15] G. Farnebnack, "Two-frame motion estimation based on polynomial expansion," in *Proc. 13th Scandinavian Co Image Analysis.*, 2003, pp. 363-370

[16] Popoola, O.P, Kejun, W, "Video-based abnormal human behaviour recognition a review*", IEEE Trans. Syst. M Cybern.* Part C. Rev. **42**, 865878 (2012)

[17] Dalal, N., Triggs, B, "Histograms of oriented gradients for human detection" *International Conference on Compu Vision and Pattern Recognition*.(2005)

[18] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb, "A large video database for human moti recognition" in *IEEE International Conference on Computer Vision*, IEEE, 2011

[19] Chen, M., Hauptmann, A, "MoSIFT recognizing human actions in surveillance videos" Computer Science. Dep. 929–9 (2009)

[20] T.-H. Yu, T.-K. Kim, and R. Cipolla, "Unconstrained monocular 3D human pose estimation by action detection a cross-modality regression forest," *in Proc. IEEE Conference on Computer Vision and Pattern Recognition,* 2013, 3642–3649.