An Approach of Understanding Human Activity Recognition and Detection for Video Surveillance using HOG Descriptor and SVM Classifier

Chandrashekar M Patil
Professor,
Department of ECE
VVCE, Mysuru
patilcm@vvce.ac.in

Jagadeesh B
Assistant Professor,
Department of ECE
VVCE, Mysuru
jagadeesh.b@vvce.ac.in

Meghana M N
Department of ECE
VVCE, Mysuru

***Abstract*—In this video surveillance moving object detection and recognition is the important research area of computer vision. Detection and recognition of moving is not easy task as continuous deformation of objects takes place during movement. Any moving objects has several attributes in temporal and spatial spaces. In spatial space object vary in size where as in temporal space it vary in moving speed. This work mainly focuses on multiple human detection and activity recognition. Multiple human video datasets are considered and in order to detect and track multiple human. Background subtraction technique is used for detecting moving multiple humans. Histogram of Oriented Gradient feature descriptor is used to extract features. For human activity recognition Support Vector Machine classifier is used.**

Keywords—Computer vision; activity detection; background subtraction; HOG descriptor; SVM classifier

I. INTRODUCTION

Computer vision is mainly used to study about how to interpret, reconstruct and understand 3D scenes from its 2D images in terms of the properties of the structures present in the scene. Computer vision mainly includes methods for acquiring, analyzing, processing and understanding digital images. Video processing is a prominent research area in the field of computer vision. . The detection and tracking of moving objects and activity recognition of these objects in video surveillance is a one of the important task.

Human detection and tracking is a major component in many of the intelligent video management and monitoring applications in recent times. This finds application in surveillance video analysis for security, sports video analysis, detection of abnormal activities, patient monitoring, traffic monitoring and many more. . Human activity recognition mainly used for human-to-human interaction as it provides information about person's identity, their personality and many more. As a result it has many applications in video surveillance systems, human-computer interaction and robotics for characterization of human behavior all these require multiple activity recognition system.

The two main approaches of detecting and tracking human are frame difference method and background modeling [2] method. Frame difference method is most suitable for no change in background and when there is relatively static situation. Background modeling method is based on Gaussian mixture model (GMM) [3], Graph cut method. GMM and Graph cut methods are more complex and large amount of calculation is involved.

The activity recognition approaches can be termed as local or global approaches. Local approach of video analysis mainly uses local interest points wherein each interest point contains a local descriptor which describe the characteristics of a point. By the analysis of these descriptor motion analysis is done. Scale Invariant Feature Transform (SIFT) and Space Time Interest Points (STIP) are some of the most commonly used local descriptors for videos. Global approach mainly uses the overall movement characteristics of the video. Most of the methods make use of optical flow to represent motion in a frame of video.

The classifier used for human activity recognition is mainly categorized into three basic types. Conditional Random Field (CRF), Hidden Markov Model (HMM) and Support vector machine (SVM) [6]. Where the CRF and HMM belongs to state model method, wherein due to the continuous change in activity sequence, activity recognition can be manipulated by modeling. Whereas, SVM use nonlinear classification function which is established by known samples to classify activities and hence overcomes the difficulties of parameter estimation in state model method. There is no need for considering the probability distribution. Hence it has its own wide variety of application.

In the present work SVM based classifier is designed to recognize multiple human activity considering two types of datasets namely, UT-interaction dataset consist of continuous activity execution which contains five human-human interactions like shake-hands, hug, push, kick and punch and the lengths of the videos are around 30 seconds. This video dataset are recorded in 2 different background setup with static background and two sets of own dataset is considered performing seven different activities like stand, walk, hug, punch, kick, handshake and fallback with duration of 30

978-1-5386-3243-7/17/$31.00 ©2017 IEEE.

seconds and both datasets are considered with static background.

II. Previous Survey

In this section an overview of latest development of human activity analysis has been shown.

Author	Year	Description
Rajvir Kaur[2]	2014	Background modelling technique is used for detecting and tracking human in video
Ahmad Jalal [3]	2014	Human tracking and activity recognition system which mainly uses body joints features for recognition
Myo Thida[4]	2013	Macroscopic and microscopic modeling technique is been used for human activity detection and tracking in crowded scenes
Mohamed Elmikaty[6]	2012	Histogram of Oriented Gradients and Shape Context based object detector are the two object detector which is used for pedestrian detection
Caroline Rougier [7]	2011	Detection of falls is been studied specifically by analyzing human shape deformation in a given video

Human activity recognition (HAR) system has wide variety of application in areas like social gathering wherein it is mainly concerned with security aspect, robotics, video surveillance and many more. Some of the issues like less accuracy for crowd analysis, slow and poor processing of compressed videos may lead for development of better algorithm. It aims in providing security for public since crime, terrorist activities etc. are increasing at a high rate.

III. Background Study

A. *Background Subtraction*

For detection moving human in a video sequence this technique is the basic and widely used step. In human detection the region of interest will be moving human in its foreground. It detects moving human by taking difference between the current image frame and reference image, here the reference image is the background image of the video take under static background condition. A simple way to implement this technique is to take a background image as a reference frame denoted by B and a frames obtained at time interval t denoted as $C(t)$. Using simple arithmetic calculations it is possible to find out the human simply by using image subtraction technique for each pixels in $C(t)$, the pixel value of current image frame is denoted by $P\,[C(t)]$ and subtract it with its corresponding pixel value at the same position of the background image which is denoted as $P\,[B]$.

The mathematical representation is given as:

$$P\,[F(t)] = P\,[C(t)] - P\,[B] \quad (1)$$

The difference image which is obtained will show some of the intensity components for the pixel locations which have changed in the two frames considered for background subtraction. This approach will work well when all the foreground pixels are moving and all the background pixels are static in nature. A threshold T is used on this difference image in order to improve the subtraction.

The mathematical form for thresholding is written as:

$$P\,[C(t)] - P\,[B] > T \quad (2)$$

Thresholding is the commonly used technique that computes a region as a set of pixels. The output obtained will be a single region wherein it represents multiple objects. The subtraction value obtained by taking difference between two images must be greater than a threshold value T, then the foreground image will be extracted. If the difference value is less then threshold T then no foreground image will be detected.

B. *Median Filtering*

Median filtering belongs to nonlinear operation, in nonlinear method thresholding and image equalization are performed. In this proposed work median filtering technique is used. This filter is mainly used to remove noise components by preserving edges. Working of median filter is carried through moving from pixel to pixel in an image and replacing each value with its median. The median filter is used to reduce noise in an image by preserving some of the useful detail present in the image. In median filtering, each pixel in an image and its nearby neighbors pixels are considered to know whether it is representative of its surroundings or not. Median is obtained by sorting all the pixel value in ascending order. The middle value of the sorted sequence is treated as median then replacing the pixel with the middle pixel value.

C. *Histogram of Oriented Gradient descriptor*

Histogram of Oriented Gradients (HOG) is used to extract feature from an image. HOG features are extracted from each binary image. The technique is used to count the occurrences of gradient in the localized portions of an image.

The algorithm of implementing HOG feature extraction is as follows:

1. The whole picture is segregated into small connected parts called as cells. HOG directions is founded for all the pixels in connected path.

2. Discretize all cell into corresponding orientation bins for $[0^0,180^0]$ periods depending on the gradient orientation and all pixels inside cell gives a weighted vote to its angular bin.

3. Grouping the obtained cells into large interlinked blocks and normalize this obtained gradient strengths which represents the histogram for each blocks.

4. The histogram obtained for each blocks represents the descriptor.

D. SVM Classifier

Support Vector Machine (SVM) is a supervised learning model which is associated with the learning algorithm and it is used to analyze the data required for classification. SVMs are based on the decision plane concept that describes decision boundaries. The decision plane separates a set of objects which belongs to different classes. It accomplishes the task of classification by creating hyperplanes in a multi-dimensional space. SVM supports classification as well as regression tasks and can handle multiple categorical and continuous variables. For categorical variables, a dummy variable is constructed with values as either 0 or 1. SVM uses iterative training algorithm to develop an optimal hyperplane, which is used to minimize an error function.

IV. METHODOLOGY

Figure 1 shows the approach of HAR. The foremost step in multiple human detection and HAR is acquiring the video datasets wherein it contains multiple human and each individual performing different activities. In this work datasets in which human activities like walk, stand, handshake, punch, hug, fallback and kick are considered. HAR starts with reading a video files. Once video is read next step is to convert this video into consecutive frames, it will extract 30 frames per sec. processed directly.

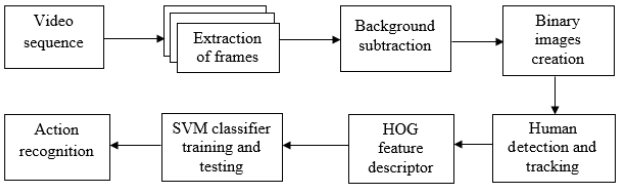

Fig. 1. *Block diagram of approach used in Human Activity Recognition*

Extraction of frames is necessary as videos cannot be processed directly. Later, background subtraction technique is used to find the moving humans. In this technique a background image is considered, where each frame is subtracted by background image to obtain foreground images which shown the moving humans location. Convert the obtained foreground RGB image to gray scale images. Onto this result 2-D median filtering is used to remove noise components present in the video.

Once after noise removal is done the gray scale images will be converted to binary images of 0s and 1s, where binary 1 is used for representing human region which is filled with white color and apart from moving human region binary 0 is used which represents absence of humans. Hence, binary image creation is useful for extracting any moving humans and objects in a video sequence. Followed by which dilation process is carried out which is typically applied on binary images obtained.

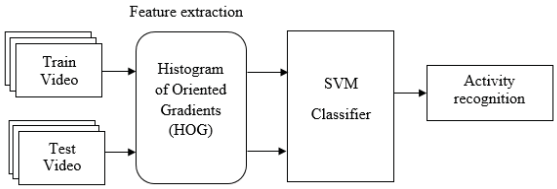

Fig. 2. *Framework of Human Activity Recognition*

Figure 2 shows Human activity recognition in which HOG feature descriptor and SVM classifier along with train and test video dataset.

Once after detecting each individual human the next stage is recognizing their activities. After detecting the moving humans in a video it is also necessary to determine the number of humans present later which activity recognition is carried out. HAR consist of two phases: training and testing.

The flow chart of training and testing phase is as shown Figure 3. In training phase the video dataset will be loaded first, then frames extraction is carried out. Training folder is created which contains the frames belonging to particular activities. Useful features are extracted for each activities being performed. HOG feature extraction technique is used for extraction of features. SVM classifier is used to train this extracted features and is saves this trained SVM model which will be further used for testing. In testing phase, the testing video is loaded and frames extraction, back- ground subtraction, binary image creation and HOG feature extraction steps will be done on the loaded test video. The obtained result will be inputted to the earlier trained SVM classifier, depending upon the feature match SVM classifier will recognize the particular activities performed.

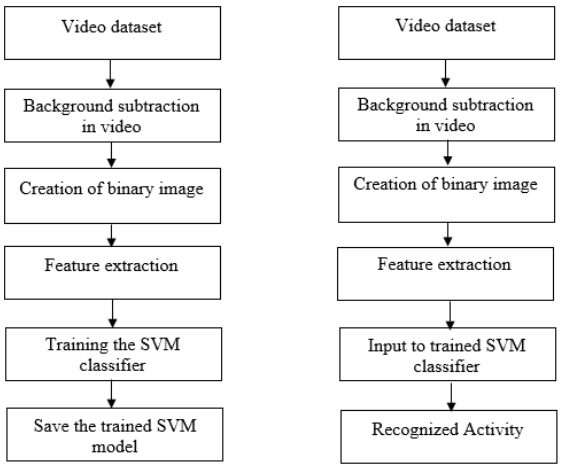

Fig. 3. *Flow chart of Training and Testing phase*

V. EXPERIMENTS AND RESULTS

This algorithm is based on UT-interaction and own dataset for human activity recognition.

A. Test video 1

UT-interaction dataset comprised of 4 people performing 5 different activities like stand, walk, handshake, punch and kick. The recorded video is of duration 13 seconds with a memory size of 1.21MB. Pixel resolution of video is 720 X 480 with frame rate of 29 frames per second. Table I indicates recognized actions categories for each frames dataset and its Figure 4 shows the recognized activities in UT-interaction.

Fig. 4. *Recognized activities in UT-interaction dataset*

B. Test video 2

Own dataset video comprised of 5 persons performing 7 different activities like walk, stand, hug, handshake, punch, kick and fall back. Video duration is 27 seconds with size 2.97MB and 720 X 480 pixel resolution. Frame rate of video is 25 frames per second. Table II indicates recognized actions categories in each frames and Figure 5 shows the recognized activities in own dataset.

Fig. 5. *Recognized activities in own dataset*

TABLE I. GENERALIZED TABULATION FOR UT-INTERACTION DATASET

Frame number	Action	Person
1	Handshake	2
2	Stand	1
	Punch	1
	Walk	2
3	Stand	1
	Kick	1
	Walk	2
4	Stand	1
	Walk	3

TABLE II. GENERALIZED TABULATION FOR OWN DATASET

Frame number	Action	Person
1	Walk	2
	Stand	1
2	Hug	2
3	Walk	1
	Handshake	2
4	Kick	1
	Fall back	1
5	Punch	1
	Fall back	1

VI. CONCLUSION AND FUTURE WORK

The proposed work gives a solution for human detection and activity recognition. Although many works have been carried out, the proposed work provides excellent results for various kinds of video datasets considered. The human detection using background subtraction for static video gives effective result and with HOG feature extraction and SVM classifier recognition of human activities provides good recognition result with less minimum number of false detections. Use of UT-interaction and own datasets achieves a higher rate of recognition. The results obtained demonstrate that the method and efficient. Therefore, the proposed technique can be regarded as a best choice for human detection and activity recognition for video surveillance application.

Future work aimed towards minimizing the false detection like shadow and other reflections of human, it can also incorporate HAR for moving background scenario and recognizing various activity like talking, eating, etc. which can be done by understanding the body pose of each individual human in the video.

REFERENCES

[1] P. Wang, J. Su, W. Li, and H. Qiao, "Adaptive visual tracking based on discriminative feature selection for mobile robot," in Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2014 IEEE 4th Annual International Conference on. IEEE, 2014, pp. 55–61.

[2] R. Kaur and S. Singh, "Background modelling, detection and tracking of human in video surveillance system," in Computational Intelligence on Power, Energy and Controls with their impact on Humanity (CIPECH), 2014 Innovative Applications of. IEEE, 2014, pp. 54–58.

[3] A. Jalal, S. Kamal, and D. Kim, "Depth map-based human activity tracking and recognition using body joints features and self-organized map," in Computing, Communication and Networking Technologies (ICCCNT), 2014 International Conference on. IEEE, 2014, pp. 1–6.

[4] M. Thida, Y. L. Yong, P. Climent-P´erez, H.-l. Eng, and P. Remagnino, "A literature review on video analytics of crowded scenes," in Intelligent Multimedia Surveillance. Springer, 2013, pp. 17–36.

[5] D. Zhang, Y. Lu, L. Hu, and H. Peng, "Multi-human tracking in crowds based on head detection and energy optimization," Information Technology Journal, vol. 12, no. 8, p. 1579, 2013.

[6] M. Elmikaty, T. Stathaki, P. Kimber, and S. Giannarou, "A novel two-level shape descriptor for pedestrian detection," 2012.

[7] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Robust video surveillance for fall detection based on human shape deformation," IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 5, pp. 611–622, 2011.

[8] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: A review," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 1, pp. 13–24, 2010.