Detection of Human Interaction

Aman Kumar, M. Poonkodi, Prateek Gupta, Ankit Kumar Gupta

Abstract— **Many works have been done in the field of activity recognition. In all those works activity was carried out by a single person. Very few amount of work has been done to recognize the interaction between two persons. This paper aims to determine the interaction activity between two persons recorded using a camera. A still camera with 360 degrees rotation has been used. Different computer vision techniques have been used to extract features from the videos. Different classification techniques such as SVM, KNN, Decision Tree, Random Forest have been applied on the extracted features. Along with individual classifiers, Ensembling techniques such as Bagging, AdaBoost have been used. All these models have been compared on the basis of accuracy. Python 3 in Anaconda environment has been used to program all these models.**

Index Terms—**Human Interaction, Flow, Machine Learning**

1) INTRODUCTION

Making the machines to think and work like a human being is the new interesting challenge in the field of engineering i.e. Computer Vision. Computer Vision is the science that aims to make machines capable of understandable, reading, analyzing, automatic extraction of data from an image or a sequence of images like a human being. It deals with how computer systems can gain high level language from images. Monitoring human activities is quite interesting and recent trend in the market. Various devices and sensors are commercialized in market for observing and monitoring the human activities such as heart rate, blood pressure, sugar level, body position and movement and various other environmental factors such as light, temperature, humidity or pollution level. The simplest or the most convenient way for a video surveillance, is to allocate a dedicated person to monitor the subject and alert if there is any danger. In this paper we aim to make the machine understandable to monitor the subject and to take the required decisions. A still camera which can rotate 360 degrees is used to monitor the subject's movement. The camera will be in ideal state when there is no movement/presence of the subject within the premises and if there is any movement it will be in active state. The captured video is converted into high quality sequences of images. System will identify various interest points from the images and monitor at regular intervals to identify the daily based activities of a human being and can make decisions in a critical situation such as falling down. It is quite useful in home care, active aging, military intelligence, prisons monitoring as a purpose of safety.

Aman Kumar *Department of CSE, SRM Institute of Science and Technology. Chennai, India, (e-mail: aman29091996@gmail.com)*
M. Poonkodi *Department of CSE, SRM Institute of Science and Technology. Chennai, India, (e-mail: poonkodi.m@vdp.srmuniv.ac.in)*
Prateek Gupta *Department of CSE, SRM Institute of Science and Technology. Chennai, India, (e-mail: guptaprateek976@gmail.com)*
Ankit Kumar Gupta *Department of CSE, SRM Institute of Science and Technology. Chennai, India, (e-mail: ankitguptand@gmail.com)*

2) RELATED WORKS

Various papers have been published on the topic video surveillance system and detection of basic human activities. Very few work has been done on detection of interaction between two persons. One such work is "Structured Learning of Human Interactions in TV Shows"[1] by Patron Perez A., Marszalek M. Video surveillance over wireless network is being used at many places such as health care, weather monitoring, tracking wild animals, cybercrime investigation etc. Recent work by Yun Ye [2] provides the detailed overview of technologies developed for wireless video surveillance system. A video surveillance system consists of video capturing, pre-processing compression and transmission and video analysis at receiving end. The main focus and the ultimate goal is to maximize the video quality received at the receiver hand for video compression and transmission over wireless network. A fixed camera is used for monitoring purpose as fixed camera always takes advantage in static background. Video is recorded at monitor site by sensor node for further processing and transmission. Various schemes have been classified such as unequal error, protection, scalable video coding, Cross layer control out of which Cross layer control seems to be the desire measure for optimal resource allocation. Received video is further processed for high quality images and privacy and security systems. The paper written by M. Valera and S.A. Velastin [3] provides with the summary of research work done till date in the field of automated visual surveillance system. The ability to recognize objects and humans and identifying their actions and interactions from the data gathered from sensors is very essential component. The main stages in this are: moving object detection, tracking and behavioral analysis. Various modular sensors have been used for monitoring purpose such as infrared, audio, video and various image processing techniques to constitute the low level part of these systems. The indeed demand of law and military applications makes automated video surveillance an essential application of Computer Vision Domain. On the other hand, paper written by Juan A. Botia [4] states the process for development of Ambient Assistant Living System which allow elderly to live alone and independently. The process covers all the major phases such as requirement analysis, architectural design etc. The major key factor of this system is that is adapts the behavior of the monitored elder and is being used as a commercial product in South-East of Spain for elders. Recent Achievements in human tracing behavior include the paper by Piotr Augustyniak [5] which provides benefits of using two mobility sensors: visual flow based image analysis and accelerometer data. It provides indoor and outdoor

recognition of elementary poses. Accelerometer and video based systems are used for dataset formation. The database consists of real life data such as recordings from the subject and using the dynamic time wrapping algorithm for measuring the distance between the different actions made by the subject. Complete focus on polar histogram based method of visual pose recognition by using the combined system of visual flow based image analysis and accelerometer data. Results obtained by the recognition of video system, accelerometer based system, combined accelerometer and video based system and by accelerometer based wearable system are 95.5%, 96.7%, 98.9%, 80% respectively.

3) DATASET

The dataset used consists of 300 videos which have been extracted from different TV shows[6]. It consists of four different interactions namely handshake, hug, highfive and kiss. There are 50 videos for each activity which accounts for 200 videos. The rest 100 videos in the dataset are negative videos. Negative videos are videos which contains activities other the four given actions. The number of frames in each video ranges from 30 to 600 frames. The clips may start from people walking towards each other or directly from the moment of interaction. This dataset has been divided into two parts- one for training and one for testing..

4) METHODOLOGY

A lot of work has been done in the field of activity recognition in which only one person performs the activity. This paper aims to recognize interactions between two people. It is assumed that both people face each other. In order to determine the interaction computer vision techniques have been applied. A general flow of the methodology is represented by the architecture diagram given in Fig 1. Points of highest intensity are referred as Interest Points. Seven interest points have been determined in each frame of the video clips. A time based flow between the interest points has been calculated.

Each video clip consists of a number of frames. Frames per second (fps) refers to the number of frames in video clip in a second. The number of frames in each video ranges from 30 to 600. Location of interest points in each frame varies. Each video clip consists of various objects along with the two persons interacting with each other. These objects are static. These form the background of the video clips. Consider two people having a handshake. Along with them different objects such as wall, chairs can be present in the frame. These objects don't contribute to the interaction being done. It is necessary to extract the persons involved in the interaction alone. Moving foreground needs to be extracted from static background. Background Subtraction based on "Improved adaptive Gaussian mixture model for background subtraction" [7] and "Efficient Adaptive Density Estimation per Image Pixel for the Task of Background Subtraction" [8]. It is a Gaussian Mixture based Background/Foreground Segmentation Algorithm [9]. It determines background and foreground on the basis of Gaussian Distribution. For each pixel in frame, Gaussian distribution is determined.

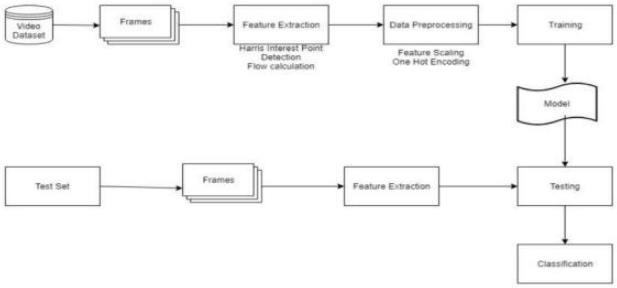

Fig 1 Architecture Diagram

Harris Corner Detector is the algorithm to detect the corner in an image or frame. A corner is also known as an interest point which gives the maximum feature of an image which can be used to identify the change in image, motion calculation in corresponding frames etc. An interest point is the intersection of two or more edges. It the simplest and reliable algorithm to calculate interest point in an image. Color to grayscale: This is the very first and important step in this algorithm. All the color images need to be converted into grayscale which is 2 Dimension. Gray Scale images are easier to process and have a very high accuracy and calculation speed. The Mathematical notation of Harris Corner is:

$$F(m, n) = \sum w(a, b)\, a,b\, [K(a + m, b + n) – K(a, b)]\, 2 \quad (1)$$

Maximization of F(m, n) is required for corner detection i.e. Taylor Function is applied to above function.

$$F(m, n) \approx [m\ n]\, T\, [\ m\ n\ ] \quad (2)$$

$$\text{Where } T= \sum a,b\, w(a, b)\, [\ IxIx\ IxIy\ IxIy\ IyIy\ ] \quad (3)$$

Harris corner detector consist of 4 parameters and are given as follows: Image :- Convert the RGB image into grayscale and float32 type. Grayscale is mandatory. Blocksize :- Neighborhood size for corner detection. Ksize :- Sobel derivative parameter K:- Parameter used in Harris equation.

The change in the position of intensity point is referred as flow. Flow between each consecutive frame is determined. Two different techniques have been applied to find out the flow between consecutive frames – Manhattan distance and Euclidean distance. Average of the flow calculated for each consecutive frame for each interest point is determined. This is referred as average flow. Average flow for each video clip is written down into a csv file along with the interaction represented in the video. Machine Learning models are applied on this csv file as dataset. Machine Learning models have been applied on the average flow value determined for each interest point. Various different classification techniques such as SVM, Decision Tree have been used. In order to obtain high accuracy values, parameter tuning has been done. Ensembling techniques such as Bagging and Boosting have been applied.

5) RESULT

In the given paper, accuracy is the parameter used to determine the efficiency of model. Different classification models have been used. Table 1 gives the accuracy obtained for different models. Table 2 gives the accuracy percentage obtained for different Ensembling techniques. Among individual classifiers SVM gives the highest accuracy of 36.48%. Parameter tuning for SVM is done to obtain high

accuracy values. On tuning the 'c', 'gamma' and 'kernel' values for SVM, an accuracy of 38.33% has been obtained for rbf kernel.

Table 1 Classification Model Accuracy

Classification Model	**Accuracy**
KNN Classifier	25%
Decision Tree	26.67%
SVM	36.48%
Naïve Bayes	30%
Random Forest	23.33%

Table 2 Ensembling Model Accuracy

Bagging	35%
AdaBoost	38.33%

6) CONCLUSION

This paper aimed to detect daily life interaction between human beings. Interaction between people is recorded using a still camera. Computer vision techniques have been used to extract features from the video clips. Different classification models such as KNN, SVM, Decision tree have been applied on the extracted features. Average flow has been calculated using Manhattan distance and Euclidean distance. Highest accuracy of 36.48% has been obtained with SVM using the Euclidean distance as the measure of flow calculation. On parameter tuning, accuracy of SVM has been raised to 38.33% using rbf kernel. Ensembling technique such as AdaBoost also gives an accuracy of 38.33%.

REFERENCES

[1] Structured Learning of Human Interactions in TV Shows Patron-Perez, A. IEEE Transactions on Pattern Analysis and Machine Intelligence.
[2] Y. Ye, S. Ci, A. K. Katsaggelos, Y. Liu, Y. Qian, "Wireless Video Surveillance: A Survey" IEEE Access, vol. 1, 2013, pp. 646-660.
[3] M. Valera and S.A. Velastin," Intelligent distributed surveillance systems: a review".
[4] Juan A. Botia, Jose T. Palma, Ana Villa, "Ambient Assisted Living system for in-home monitoring of healthy independent elders", Expert Systems with Applications, vol. 39, 2012, pp. 8136-8148.
[5] Piotr Augustyniak, Magdalena Smolen, Zbigniew Mikrut and Eliasz Kantoch, "Seamless Tracing of Human Behavior Using Complementary Wearable and House-Embedded Sensors".
[6] High Five: Recognising human interactions in TV shows Patron-Perez, A., Marszalek, M., Zisserman, A. and Reid, I. Proceedings of the British Machine Vision Conference (BMVC), Aberystwyth, UK, 2010.
[7] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction", International conference on Pattern Recognition 2004.
[8] Z. Zivkovic, "Efficient Adaptive Density Estimation per Image Pixel for the Task of Background Subtraction", Pattern Recognition Letters, volume 27, 2006.
[9] https://docs.opencv.org/3.4/db/d5c/tutorial_py_bg_subtraction.html

Aman Kumar is currently pursuing B.Tech in Computer Science and Engineering from SRM Institute of Science and Technology, Chennai. He will graduate in May 2019. He has published a paper in Innovations in Soft Computing and information Technology, Volume 3.

M. Poonkodi Has received M.Tech Degree from the department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India, in the year 2011. She is currently an Assistant Professor in the department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. Presently being a research scholar in the same institution, her research interests include computer vision, pattern recognition, video analytics.

Prateek Gupta is currently pursuing B.Tech in Computer Science and Engineering from SRM Institute of Science and Technology, Chennai. He will graduate in May 2019.

Ankit Kumar Gupta is currently pursuing B.Tech in Computer Science and Engineering from SRM Institute of Science and Technology, Chennai. He will graduate in May 2019.