A Review Paper on Human Detection & Tracking using Background Subtraction Modelling

Nisha Thakur[1] Mr. Sachin Meshram[2]
[1]M.Tech Scholar[2]Assistant Professor
[1,2]Department of Electronics & Telecommunication Engineering
[1,2]CEC, Bilaspur, Chhattisgarh, India

Abstract**—***In computer vision community, 'Action' and 'Activity' are frequently interchangeable terms. In this thesis, 'Action' is defined as simple motion patterns which are normally exhibited by a single person for a very short duration. Examples of action include running, walking and waving. On the other hand, 'Activity' is defined as a complex sequenceof actions where several humans are involved and they interact with each other in a constrained manner. Activities are typically characterized by much longer temporal durations. Examples of activities are two persons shaking hands, a football team scoring a goal and a coordinated bank robbery multiple persons. Recognition of human actions from multiple views needs high level interactions with wide applications such as video surveillance, human computer interaction, motion analysis, video indexing and sports video analysis. Various researches have been conducted to develop a fully automated video surveillance system that can mimic the activity of human brain in identifying events in moving objects. However, due to the advanced technological breakthroughs in video capturing devices, increase in the number of cameras used, dynamic nature of the objects (human, vehicles, animals) in the video has made this process a very challenging field of research. The main goal of this research work is to propose a video surveillance system that is robust and can perform Human detection, tracking and classification and can be used to provide security protection to both private and commercial sections.*

Keywords—Background Subtraction, Image Segmentation

I. INTRODUCTION

Vision is the most superior of human senses and it is no surprise that images and videos convey more important information during human perception. Image processing paved a wide spectrum of application fields due to their variety of light energy, namely visible, Ultra violet, X-rays, Gamma rays, Infrared, Microwaves and radio waves. In general, image processing is distinguished as low level, mid-level and high level processing. The lower level processing, also termed as image preprocessing, involves primitive operations on images such as noise removal, contrast enhancement and image sharpening. The mid-level or intermediate processing on images involves tasks such as segmentation, object representation, description and classification. The higher level processing involves image recognition, image understanding or computer vision.

A. Action Description

The research in the fields of computer vision has dealt with the analysis of temporal sequence of image data. The huge availability of low cost and high quality digital cameras with massive storage capabilities, and high speed bandwidths has all contributed to realize the role of video processing in various application domains. Also, human biological visual systems are well managed to handle spatiotemporal information, which is yet another reason for the popularity of video data analysis. Video based tracking and analysis helps to observe the motion patterns in a non-intrusive way. This task has to be well modeled when dealt with human subjects. Video surveillance systems play a very important role in the circumstances where continuous observation by visual analysts is not possible. Normally, visual analysts continuously monitor and collect data from various cameras, and report to the authorities when necessity arises. But, it is manpower intensive. Hence, it is necessary to build automatic human action recognition system and build a higher level behavior modeling for the events occurring in the scene.

In computer vision community, 'Action' and 'Activity' are frequently interchangeable terms. In this thesis, 'Action' is defined as simple motion patterns which are normally exhibited by a single person for a very short duration. Examples of action include running, walking and waving. On the other hand, 'Activity' is defined as a complex sequence of actions where several humans are involved and they interact with each other in a constrained manner. Activities are typically characterized by much longer temporal durations. Examples of activities are two persons shaking hands, a football team scoring a goal and a coordinated bank robbery multiple persons. Recognition of human actions from multiple views needs high level interactions with wide applications such as video surveillance, human computer interaction, motion analysis, video indexing and sports video analysis.

B. Human Motion Analysis

The conventional Human Motion Analysis (HMA) system is able to detect, track and identify the moving humans in a video sequence and may include recognizing their action. The types of interaction of HMA system with the environment are as follows:

1) Passive:

It simply captures and stores the visual information in an organized fashion without performing any analysis.

2) Active:

It controls and adjusts the acquisition device parameters, namely pan, zoom and tilt effects, depending on the external environment conditions.

The HMA typically has the following three basic steps:

1) Detection: It involves finding the answer, 'Is there motion (corresponding to a human) present in the scene?' and essentially requires low level processing of images.
2) Tracking: It answers the question: 'Where is the human moving?'. The tracking is of major importance to HMA. It needs some kind of history to be maintained for the purpose of action recognition and it involves the

mid-level processing on the history of images. However, sometimes there may be a considerable overlap between detection and tracking algorithms.

3) Action Recognition / Behavior Understanding: It is a high level vision step, which involves interpreting the information derived in the above mentioned steps in order to answer the question 'What is the human doing.

A schematic representation of the three phases of a HMA system is shown in Figure 1.1.

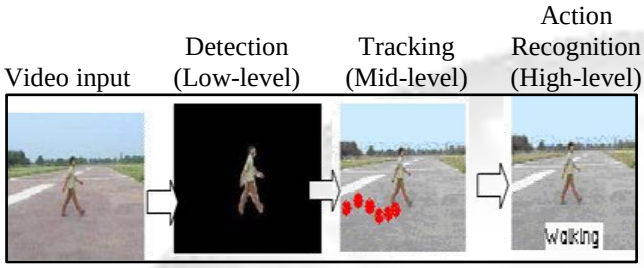

Fig. 1.1: Three Phase of HMA System

1) Applications of HMA:

The remarkable prospective in research on HMA becomes evident when looking at the applications that may be benefited from the research works in this domain. Some application areas that highlight the potential impact of vision based action recognition system are listed in Table 1.1 and discussed in this section

a) Kinesiology:

Kinesiology, the study of biomechanics, involves development of human body models aiming to improve the efficiency of human movement. This is widely used in medical studies of orthopedic patients and such studies need detailed information about movement of body parts and joints this information is gathered in an intrusive way by placing retro-reflective markers or light emitting diodes (LED) on the human body.

b) Computer Graphics and Animation:

HMA can be used to develop a high level description of movements of dance and sports field. This has been used to study and synthesize realistic motion patterns of virtual world humans

c) Behavioral Biometrics:

Biometrics involves the study of approaches and algorithms for uniquely recognizing humans, based on human physical and behavioral cues. The problems of gait recognition and gesture recognition are solved by the appearance based tracking approaches. The physical body parts are marked and their motion trajectories are mapped to identify the human.

d) Content based Video Analysis:

Since internet service providers (ISP) face persistent growth, it has become compulsory to develop efficient indexing and storage schemes to improve the user feedbacks for content based video searching. This involves the learning of flow patterns from raw video and summarizing the content of video. The enhanced content based video summarization with corresponding advances in content based image retrieval. The common approach for indexing and retrieval is to query the video database using semantic action descriptors like 'videos where person kicks somebody'. So, the retrieval is the kind of process to deliver the analogous videos according to action based features in the query video.

e) Surveillance and Security:

Video surveillance and related security systems mostly rely on multiple video cameras, monitored by a human operator who should know the activity of interest in the field of view. The monitoring efficiency and accuracy of the human operators can be stretched for increasing the number of cameras and their deployments. Hence, security agencies are seeking the vision based solutions which can replace or assist a human operator. Generally, the existing surveillance techniques performed a post operation task and detected the abnormal events based on the actions that have been committed. These techniques, many times, need manual intervention for real time detection. However, activity detection systems incorporated with 'Smart Surveillance Systems' (SSS) in real time detection are also possible without manual intervention, and are used in banks, ATM centres, parking lots, supermarkets and department stores. So, the conflicts in such surveillance applications are the main factors of SSS which affect the privacy of human beings.

f) Human Computer Interface:

Understanding the interaction between a computer and a human remains one of the enduring challenges in designing human-computer interfaces. The visual communication is the most important mode of non-verbal communication between human and computer. Effective utilization of visual cue features is important to improve interaction of computer with humans.

C. Properties of Shape Features

The machine vision has super excellent performance towards complicated problem solving with small complexity. In addition, choosing appropriate features for a shape recognition system must consider the kind of features suitable for the task. The techniques to describe the shape of a deformable object have been extensively studied by researchers for the past few decades. Mingqiang et al (2008) compared many shape representation techniques to characterize a deformable object. These techniques belong to the following three main categories:

1) Boundary based and region based methods
2) Spatial domain and frequency domain methods
3) Information preserving and non-information preserving methods

In boundary based shape description, the shape boundary points are only used to represent the shape. But in the region based methods, the combination of boundary and interior points are used to represent the shape. In the spatial domain methods, the point feature basis helps to compare two shapes and the vector basis is used in frequency domain methods. The information preserving method provides an accurate shape reconstruction from the shape descriptors and the non-information preserving methods offer only partial reconstruction with a compromise on the identifiable property of shape features. The identified shape features should have the following essential properties irrespective of the description strategy:

1) Identifiable: The shapes which are similar to human perception should have closest feature correspondence.

2) Spatial invariance: The spatial transformations such as translation, scaling and rotation should not affect the extracted features.
3) Noise resistance: Features must be as robust as possible against noise, i.e., the noise strength of specific range should not affect the pattern.
4) Occultation invariance: The original shape features should not be affected during occultation.
5) Statistically independent: Two distinct features must be statistically independent and must ensure compactness of the representation.
6) Reliable: To ensure minimum intra class variability.The widely used methods for human shape representation in HMA are as follows:
7) Bounding Box - This is the smallest rectangle that contains every point in the object shape. The shape can either have a global bounding box or a collection of bounding boxes which describe the human body parts, namely head, torso, legs, hands and feet. However, this representation conveys the coarse features only.
8) Moments - The magnitude of a set of orthogonal complex moments of the image known as Zernike moments is most suitable for shape similarity based image retrieval in terms of compact representation, robustness and retrieval performance (Kontanzad and Hong 1990). But, the presence of many factorial terms of Zernike moments leads to complex
9) Chain code - the complete boundary description of the shape that ensures retaining of the shape. The chain code cannot tolerate scale and rotation variations which are common problems that occur in object tracking.
10) Fourier descriptors – The high frequency components of Fourier descriptor are associated with corners and low frequencies are associated with shape's border. The Fourier descriptor is used to transform a set of boundary pixels from the spatial domain representation into the frequency domain. The Fourier descriptors are also invariant to scale rotation and starting point of the shapes.

So, the accuracy of any pattern recognition system is based on the selection of appropriate shape featur

Background subtraction is a simple solution in motion segmentation. A static image without any object of interest (OOI) would be considered as background image. The difference in pixel level between the successive frames and the background image provides motion information. Feature extraction, in accordance with human perception, is a very complex task in shape recognition. The selection of appropriate shape feature would be considered for improving the accuracy of any pattern recognition system and related experimental analysis could be used to measure the suitability of the feature to obtain proper outcome. Human posture refers to the arrangement of the body and its limbs. This is one of the key aspects of analyzing human behavior. Likewise, implementation of action recognition system meets many challenges at each subsystem. The background subtraction suffers from clutter, illumination changes, and camera movements, and existence of either noise or partial occlusions during tracking affects feature extraction stage.

II. PROBLEM STATEMENT

Many researchers have contributed innovative algorithms and approaches in the area of human action recognition system and have conducted experiments on individual data sets by considering accuracy and computation. In spite of their efforts, this field requires high accuracy with less computational complexity. The existing techniques are inadequate in accuracy due to assumptions regarding clothing style, view angle and environment.

Hence, the main objective of this thesis is to develop an efficient multi-view based human action recognition system using shape features. During the development phase, the following two objectives have been conceived in the proposed approach:

1) Primary Objective – to develop an efficient human action recognition system using multiple views.
2) Secondary Objective – to understand human behavior model using probabilistic action graph.

A. Problem Statement of Primary Objective

The existing boundary based features are insufficient to represent the shape information due to high dimensionality and computational complexity. To solve this problem, problem statement 1 is formulated.

B. Problem Statement 1

To propose a simple and suitable scheme for extracting boundary based shape features to obtain robustness against occlusion and noise.

The solution to this problem is attained by combining the novel triangulated shape orientation context based shape features and centroid orientation context based shape features.

III. PROPOSED METHODOLOGY

The detailed literature review and the methodology implemented in this research work are discussed in the last chapter. The segmentation of human body from the video frames is a very significant process for human tracking, human focusing, and activity analysis in the field of video surveillance system. The variation in the threshold values due to lighting conditions with camera noise and the human body segmentation from video frame provide inaccurate background subtraction and it requires more computation time. This chapter proposes background subtraction techniques which uses Automatic Threshold Update (ATU) and Discrete Wavelet Transform (DWT) along with Frame Differencing (FD) method for the human body segmentation from the video frames.

A. Proposed Approach

The background subtraction is a method which uses to segment the human body from the background of an image. It is used to focus the desired human body in the frame for better understanding. It is useful in reducing the computational complexity since it is considered only the foreground objects. It is more important because the process like human body parts identification and pose modelling are more dependent on it. It is used to detect the moving regions by subtracting pixel by pixel basis from the background image. Here the threshold is fixed such that the foreground pixels are extracted from the background image. When the

pixel difference is above the threshold value, it is considered as a foreground image.

Initially, the videos are acquired by the video camera from the indoor environment. Then, the pre-processing operation is performed to enhance the quality of frames. After the pre-processing, the background is subtracted using the proposed approaches. Here, the simple Frame Difference (FD) method is considered with ATU and DWT approaches. For a video, the threshold value is very important in separating the foreground image from the background image. The threshold values are dynamic due to the practical factors which affects the segmentation process. The threshold value has to be updated to get an efficient background subtraction.

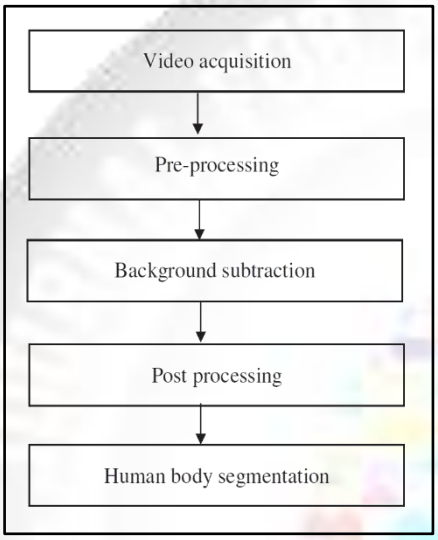

Fig. 4.1: Overview of the proposed work for the human body segmentation

B. Frame Difference Method

The gray scale and RGB based computations are proposed for this implementation. Methods 1, 2, 3 and 7 are performed based on Gray scale computation whereas RGB computation is performed in the methods. The frame difference method is achieved using pixel to pixel subtraction of the current frame from the background. When the pixel difference is above the threshold value, it is considered as foreground image. Otherwise it is known as background image. Here, the threshold value is fixed. By the experimentation with various videos, the threshold value is set as 55. The flow chart of the frame differencing method for gray scale approach is shown in Figure 4.2.

In a similar way, the RGB computation is achieved for the frame difference method except that each pixel is separated into three components namely Red (R), Green (G), and Blue (B). Each of these three components are extracted from background frame and current frame and compared with the pre-defined threshold values. The frame differencing algorithm for RGB method include, Frame Differencing Algorithm (for RGB)

- Step 0: Read the video sequences using monocular camera.
- Step 1:Set the background image which acts as reference frame.
- Step 2:Separate R, G, and B components individually for easy computation.
- Step 3:Read the current frame from the video sequence.
- Step 4:Separate R, G, and B components individually for easy computation.
- Step 5:Subtract the corresponding colour components of the background and current images.
- Step 6:Check the threshold values of colour components.
- Step 7: Display the foreground image.
- Step 8:Repeat steps 0-8, till the last frame of the video sequences.

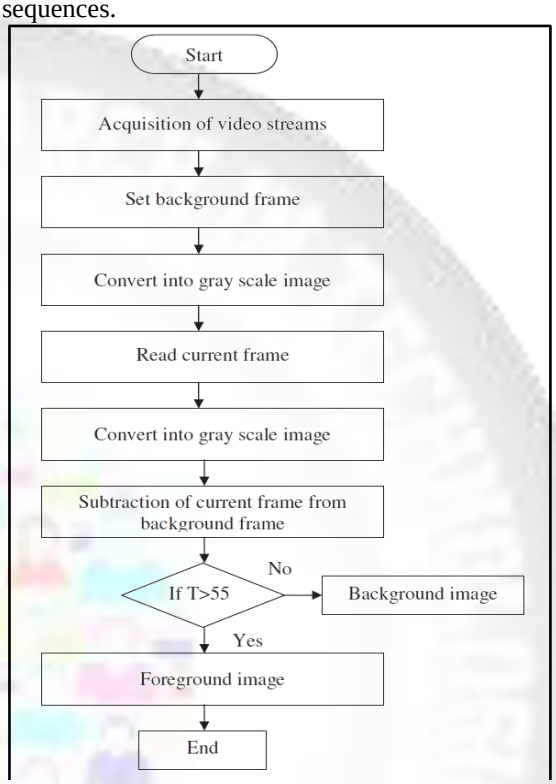

Fig. 4.2: Flow chart of frame differencing method (Gray scale approach)

IV. EXPECTED RESULT

Video surveillance systems are currently undergoing a transition from traditional analog solutions to digital solutions. These systems are increasingly becoming a part of daily life and are used in airports, banks, offices, hospitals, traffic points and even at residential apartments. They are used extensively in security conscious fields like military, intelligent traffic monitoring, legal departments and also provide high level security in border monitoring and transport environments.

Various researches have been conducted to develop a fully automated video surveillance system that can mimic the activity of human brain in identifying events in moving objects. However, due to the advanced technological breakthroughs in video capturing devices, increase in the number of cameras used, dynamic nature of the objects (human, vehicles, animals) in the video has made this process a very challenging field of research. The main goal of this research work is to propose a video surveillance system that is robust and can perform object detection, tracking and classification and can be used to provide security protection to both private and commercial sections.

REFERENCES

[1] Tom Wilson, Michael Glatz, and Michael Hödlmoser, "Pedestrian Detection Implemented on a Fixed-Point Parallel Architecture," The 13th IEEEInternational Symposium on Consumer Electronics (ISCE2009), pp. 47-51, 2009.

[2] Damien Simonnet and Sergio A Velastin, "Pedestrian detection based on Adaboost algorithm with a pseudo-calibrated camera," IEEE Image Procesing Theory, Tools and Applications, 2010.

[3] Qixiang Ye, Jianbin Jiao, Baochang Zhang, "Fast pedestrian detection with multi-scale orientation features and two-stage classifiers," Proceedings of 2010 IEEE 17th International Conference on Image Processing, , pp. 881-884, September 2010, Hong Kong.

[4] A´ kosUtasi and CsabaBenedek , "A Bayesian Approach on People Localization in Multicamera Systems," IEEE transactions on circuits and systems for video technology, vol. 23, no. 1, pp. 105-115, January 2013.

[5] BerkinBilgic, Berthold K.P. Horn, and Ichiro Masaki, "Fast Human Detection with Cascaded Ensembles on the GPU," 2010 IEEE Intelligent Vehicles Symposium, University of California, San Diego, CA, USA, pp. 325-332, June 2010.

[6] Paul Viola, Michael Jones, and Daniel Snow, "Detecting Pedestrians using Patterns of Motion nd Appearance," International Journal of ComputerVision, vol. 63(2), pp. 153–161, 2005.

[7] Stephen Krotosky, Mohan Manubhai Trivedi, "On Color-, Infrared-, and Multimodal-Stereo Approaches to Pedestrian Detection," IEEE transactions on Intelligent Transportation Systems, vol. 8 (4), pp. 619-630, 2007.

[8] BoWu and Ram Nevatia, "Detection and Segmentation of Multiple, Partially Occluded Objects by Grouping, Merging, Assigning Part Detection Responses," International Journal of Computer Vision, vo. 82, pp. 185-204, 2009.

[9] Wen-Chang Cheng and Ding-MaoJhan, "A self-constructing cascade classifier with AdaBoost and SVM for pedestrian detection," Engineering Applications of Artificial Intelligence, vol. 26, pp. 1016–1028, 2013.

[10] Rodrigo Verschae, Javier Ruiz-del-Solar and Mauricio Correa, "A unified learning framework for object detection and classification using nested cascades of boosted classifiers," Machine Vision and Applications, vol. 19, pp. 85–103, 2008.

[11] JooKooi Tan, KazukiInumaru, Seiji Ishikawa, and Takashi Morie, "Automatic detection of pedestrians from stereo camera images," Artif Life Robotics, vol. 15, pp. 459–463, 2010.

[12] GianlucaAntonini, Santiago Venegas Martinez, Michel Bierlaire, and Jean Philippe Thiran, "Behavioral Priors for Detection and Tracking of Pedestrians in Video Sequences," International Journal of Computer Vision, vol. 69(2), pp. 159–180, 2006.

[13] AzraHabibovic, Emma Tivesten, Nobuyuki Uchida, Jonas Bärgman, and Mikael LjungAusta, "Driver behavior in car-to-pedestrian incidents: anapplication of the Driving Reliability and Error Analysis Method (DREAM)," Accident Analysis and Prevention, vol. 50, pp. 554– 565, 2013.