Proceedings of the 10th INDIACom; INDIACom-2016; IEEE Conference ID: 37465
2016 3rd International Conference on "Computing for Sustainable Global Development", 16th - 18th March, 2016
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA)

A Study On Human Activity Recognition From Video

Debashis Barman
Department of CSE & IT
School of Technology, Assam Don Bosco University
Guwahati, Assam, INDIA
deb.dbuniversity@gmail.com

Usha Mary Sharma
Department of CSE & IT
School of Technology, Assam Don Bosco University
Guwahati, Assam, INDIA
ushamary.sharma@gmail.com

***Abstract*—Analysis of human activities from video is currently one of the ongoing research areas in computer science. Recognition of human activity from video has gained lot of attention because of its increasing demand in many real life applications, for e.g. video surveillance, entertainment, healthcare, child and old age homes, etc. In this paper, several steps involved in automatic human activity recognition systems, such as segmentation, tracking of motion, pose estimation and recognition of activities are studied. Common segmentation techniques such as background subtraction and Gaussian Mixture Model (GMM) are also discussed. Two different approaches for tracking of motion in video, i.e. representation and localization of the target and filtering and data association are also discussed. Finally some of the commonly available datasets used in automatic analysis of human activities from video are also mentioned in detail.**

Keywords—human activity detection; segmentation; tracking; pose estimation; activity recognition, background subtraction, GMM

I. INTRODUCTION

In recent time, automatic recognition of human activity from video has drawn attention because of its growing need in various real life environments, such as security surveillance, entertainment and gaming, healthcare, child care, academic institutions etc. Generally in a surveillance environment, this kind of technology can be used to alert the security persons or related authorities of crime or dangerous activities, for e.g. automatic recognition of a person with a bag loitering inside the station or airport. In entertainment and gaming environment, similar technology can increase human and computer interaction. Moreover, in child care, old ages home, automatic recognition of human actions may help to facilitate the entire process of care or rehabilitation. In general, the recognition of human activity from video involves different steps, such as, acquisition of input video and extraction of the frames, segmentation of human body silhouettes, motion tracking, pose estimation and finally, the recognition of the activity.

Generally, human activity recognition from video starts with the acquisition of the input video and preprocessing such as frame extraction. Segmentation is the process of dividing the image into disjoint segments such that all together the segmented parts form the original image. The segmentation step is followed by the motion tracking in which the basic idea is to detect the moving objects in the frame. Finally, various movements taken by a pose during a particular activity are estimated and activity recognition algorithms are used to identify and investigate the actions.

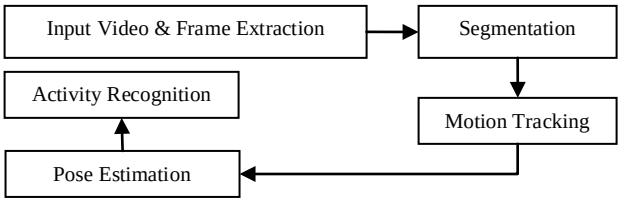

Fig. 1. Steps in Human Activity Recognition

This paper studies the various tasks associated with the human activity recognition process. The rest of the paper is organized as follows : section II briefly presents a review of the related works, section III discusses about segmentation, section IV discusses about the motion tracking process, section V discusses about the post estimation, section VI discusses about activity recognition algorithms and section VII discusses some commonly available dataset. Finally the section VIII concludes the paper.

II. RELATED WORKS

Chen et al. [1] has studied various difficulties in extraction of silhouettes and tracking of human in a real world scenario where the background model happens to dynamic and complex. Different features are extracted from the image segments, gather the feature details over a period and fuse these high level information with low level features to build a background model that varies with time. A fuzzy decision model is developed to separate the objects in motion in the foreground from the human body. The experimental results show the algorithm is very efficient as well as robust in nature. However, from results it can be deduced that the basic morphological operation, such as "cleaning" is not enough to completely recover the process of degradation for some image sequences, mainly the images in greater motion and larger objects in the

 ISSN 0973-7529; ISBN 978-93-80544-20-5

scene. In the work of Scheer et al. [2], silhouettes are extracted in a YUV colour space which is able to detect shadow. The proposed method is capable for real-time video processing without any need of transforming the colour space. This algorithm does not work well if the target object is not moving or other moving objects are present in the scene.

Setiawan et. al. [3] proposed Gaussian Mixture Model (GMM) for foreground segmentation along with an Improved HLS colour space model that is capable of differentiating shadow region from objects. The Improved HLS colour space model serves as the base description for image as it is advantageous over RGB colour space to recognize shadow portions from the object by utilizing luminance and saturation information directly and without any need of calculation of chrominance and luminance. However, Wei Niu, *et al.* [4] has proposed a real-time algorithm which is applicable for analyzing distant outdoor human activities. The effectiveness in activity analysis can be further improved by allowing intelligent control and fail-over approaches, built on top of low-level motion detection approaches such as frame differencing and feature correlation. Furthermore, Neil Robertson et al. [5] achieved action recognition via probabilistic search of image feature databases which represents previously seen actions. The Hidden Markov Models (HMM) are used to smooth sequences of actions in the video. High- level activity recognition is achieved by calculating the possibilities that the set of predefined Hidden Markov Models explains the current sequence of action.

III. SEGMENTATION

In a video analysis application where analysis of human activities is carried out, the very first step is the detection of human movements. Detection of human provides a focus of interest for the future tasks. Segmentation is one of the basic steps in image processing. Segmentation is the process of dividing the image into dissimilar portions so that all together these segments results in the original image. There are various methods available for detection of human movements in videos. Some of the commonly used segmentation methods are discussed below.

A. Background Subtraction

The most commonly used segmentation technique is called the background subtraction method. Basically the regions of interest in the image are the foreground objects of the image. The basic idea in background subtraction is the detection of the moving objects from the difference in between the current frame or the current image and a reference frame, or the background image, or the background model. Background subtraction is basically performed when the target image is a part of the video. This mechanism provides important detailsfor numerous applications in computer vision, such tracking of motion or estimation of various human poses. However, this approach is mainly based on a static background model which may not be always applicable in real life scenarios. Basically, the background subtraction algorithm is quite simple and efficient, but sometimes its simplicity may cause the inaccurate classification of the pixels. In fact, continuous updating of the background model in response to gradual changes of background may also sometimes cause difficulties.

B. Gaussian Mixture Model

Gaussian mixture model is used to deal with different background scenarios in which instead of only one-Gaussian per pixel in the background model, the pixel values at location (x,y) can be used as a mixture of Gaussians in the background model. Gaussian mixture model (GMM) can be used to overcome the problems in the above mentioned background subtraction technique. The Gaussian mixture model (GMM) has been implemented in the multi-model environments. This method develops a background model by a mixture of *K* Gaussian distributions. Normally, the value of *K* is restricted between 3 and 5 due to increase in computational complexity. The pixels in the current image are compared to the background model with every single Gaussian in the model until a match is found and if there is no match then a Gaussian with a new mean value equivalent to the current pixel color is introduced into the mixture with the initial variance. In Gaussian mixture model, each pixel in each frame is the mixture of K Gaussian distributions [11] and the probability of the pixel x_t at time t is obtained as

$$P(\mathrm{x_t}) = \sum_{i=0}^{n} w_i \eta(x_t, \Theta_i) \rightarrow (1)$$

where $w_i$weight parameter of K Gaussian factor and $\eta(x_t, \Theta_i)$ is the normal distribution of K. The normal distribution is given by,

$$\eta(x_t, \Theta_i) = \eta(\mathrm{x_t}, \mu_\mathrm{k}, \Sigma_\mathrm{k}) \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_\mathrm{k}|^{\frac{1}{2}}} \exp(Z) \rightarrow (2)$$

where μ_k is the mean and Σ_k is the covariance of K factor and Z is found by

$$Z = \left[-\frac{1}{2} (x - \mu_\mathrm{k})^T \sum_k^{-1} (x - \mu_\mathrm{k}) \right] \rightarrow (3)$$

K is the total number of estimated distribution estimated as per wi while T is the threshold value and value of T is used to be 0.2 to model the background.

IV. MOTION TRACKING

Tracking is carried out to locate the moving object or multiple objects over a time period with the help of a camera ormultiple cameras. It has a numerous applications, for example, security surveillance, human and computer interaction, traffic monitoring, video editing, video communication, medical imaging, augmented reality, etc. Motion tracking is generally a process that consumes time because of the amount of data contained in video. Furthermore, the possible need of use of object recognition techniques in the motion tracking is itself a challenging problem.

Motion tracking algorithms analyze the sequence of frames in the video and gives the movement of target objects between these frames as outputs. There are different types of motion tracking algorithms. Each algorithm has its own strengths and weakness. Therefore it is important to consider the objective of

motion tracking in choosing proper algorithm. There are two main parts of a motion tracking i.e. the representation as well as localization of the target and the filtering as well as data association.

A. Representation as well as Localization of the target

Representation as well as localization of the target is a bottom up approach. Representation as well as localization methods yield different techniques for identification of the object in motion. Localization and tracking of the object in motion depends on the algorithm used. There are some general algorithms for target representation as well as localization, such as—

1) Kernel Based Tracking or Mean Shift Tracking: Kernel based tracking or mean shift tracking is the localization technique based on maximization of a similarity measures. [6]

2) Contour Tracking: Contour tracking techniques iteratively develop an initial contour from the previous image to the new position in the current image. This method depends on using lines calculated by a optical flow based on gradient and edge detector. [7]

B. Filtering as well as Data Association

Filtering as well as data association is mainly a top-down approach, which involves integrating some initial details about the environment or the objects, dealing with objects in movements and evaluating further hypotheses. Filtering and data association techniques facilitate the tracking of complex objects, such as tracking of moving objects behind the obstacles. However, the complexity will increase if the tracker is not positioned on a fixed base but positioned on a moving base. In that case, generally an inactivity measurement is used to balance the tracker and thereby to reduce the movement of the camera [8]. Usually the complexity for these kind of algorithms is found to be higher. Two commonly used filtering techniques are discussed below—

1) Kalman Filter: Kalman Filter is a process which uses a sequence of measurements over a period that contains the inaccuracies, for e.g. noise, and gives the approximation of unknown variables which tend to be more accurate than those based on one measurement. [9]

2) Particle Filter: Particle filter is used for sampling the state-space distribution of nonlinear as well as non-Gaussian processes. [10]

V. ESTIMATION OF POSE

The main goal of the pose estimation is to determine the type of movement taken by a pose in a particular activity. Estimation can be performed using supervised and unsupervised learning techniques. Supervised learning techniques compare the unknown poses with some preset poses which are already estimated in the training. In case of unknown poses, various patterns involved in a pose cannot be known in advance. Therefore unsupervised learning approach e.g. self-learning,self-organizing methods, etc. are used.

VI. RECOGNITION OF ACTIVITY

Once the pose estimation is done, the next step is the recognition of various activities in the video. The main objective of this step is to identify the actions. It also includes different goals such as differentiating regular and irregular activities, determining different activities, etc. Holistic recognition approaches try to recognize a person in the scene, gender, etc. Recognition based on the related body parts involved in a particular action is more advantageous than considering the entire body.

VII. COMMON DATASET

In the past few years various video datasets are made available publicly for executing different tasks involved in the activity recognition process, such as the segmentation, motion tracking and object recognition, action analysis, etc. Use of these already available datasets saves both time and resources since there is no necessity of creating the video sequences from scratch. Some of the commonly used datasets are KTH, WEIZMANN, UT-Interaction, MSR action, etc [12].

The KTH dataset [13] contains six types of human actions i.e. walking, running, boxing, jogging, hand waving and hand clapping. The actions are demonstrated many times by 25 different subjects in four different scenes. All sequences were taken over similar backgrounds with the help of a static camera with frame rate of *25*fps in a resolution of 160x120 pixels.

The WEIZMANN dataset [14] contains videos 10 video sequences of jumping, running, walking, bending, sideways galloping, one-hand and two hands waving, jumping in place, jumping jack and skipping. All of the videos are recorded with the help of a static camera in a resolution of 180x144 pixels.

The UT-Interaction dataset [15] includes video sequences of continuous demonstrations of six classes of human and human interactions i.e., shaking hands, pointing, hugging, pushing, kicking and punching. There are 20 video sequences of around 1 minute length. These videos are taken with the resolution of 720x480 pixels at frame rate of 30fps. The height of the subject in each video is about 200 pixels.

MSR Action dataset [16] has 16 different videos containing all together 63 actions out of which 24 are hand waving, 14 are hand clapping, and remaining 25 are boxing. All the actions are performed by 10 different individuals. Each video sequence has different types of actions. Each video sequence is of resolution 320 x 240 with frame rate of 15 fps of 32 to 76 seconds length.

VIII. CONCLUSION

In this paper, we have presented a study of various steps as well as methods involved in the analysis of human activity recognition from video inputs which has an constantly growing demand in the computer vision related applications. In our consideration, instead of extracting all the frames from the input video, extracting the frames where some action takes place can be much beneficial. Furthermore, recognition based on the related body parts involved in a particular action is more advantageous than considering the entire body.

REFERENCES

[1]. Chen, Xia, et al. "Adaptive silouette extraction and human tracking in complex and dynamic environments." Image Processing, 2006 IEEE International Conference on. IEEE, 2006.
[2]. Schreer, Oliver, et al. "Fast and robust shadow detection in videoconference applications." Video/Image Processing and Multimedia Communications 4th EURASIP-IEEE Region 8 International Symposium on VIPromCom. IEEE, 2002.
[3]. Setiawan, Nurul Arif, et al. "Gaussian mixture model in improved hls color space for human silhouette extraction." Advances in Artificial Reality and Tele-Existence. Springer Berlin Heidelberg, 2006. 732-741.
[4]. Ke, Shian-Ru, et al. "A review on video-based human activity recognition." Computers 2.2 (2013): 88-131.
[5]. Robertson, Neil, and Ian Reid. "A General Method For Human Activity Recognition In Video." Computer Vision and Image Understanding 104.2 (2006): 232-248.
[6]. Comaniciu, Dorin, Visvanathan Ramesh, and Peter Meer. "Real-time tracking of non-rigid objects using mean shift." Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on. Vol. 2. IEEE, 2000.
[7]. Yokoyama, Masayuki, and Tomaso Poggio. "A Contour-based moving object detection and tracking." Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on. IEEE, 2005.
[8]. Black, James, Tim Ellis, and Paul Rosin. "A novel method for video tracking performance evaluation." Proceedings of the IEEE InternationalWorkshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS 03) (2003): 125-132.
[9]. Arulampalam, M. Sanjeev, et al. "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking." Signal Processing, IEEE Transactions on 50.2 (2002): 174-188.
[10]. Rincón, JesÚs Martínez Del, et al. "Tracking human position and lower body parts using Kalman and particle filters constrained by human biomechanics." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 41.1 (2011): 26-37.
[11]. Bouwmans, Thierry, Fida El Baf, and Bertrand Vachon. "Background modeling using mixture of gaussians for foreground detection-a survey." Recent Patents on Computer Science 1.3 (2008): 219-237.
[12]. Human Activity Video Datasets, https://www.cs.utexas.edu/~chaoyeh/web_action_data/dataset_list.html
[13]. KTH Dataset, http://www.nada.kth.se/cvap/actions/
[14]. WEIZMANN Dataset, http://www.wisdom.weizmann.ac.il/%7Evision/SpaceTimeActions.html
[15]. UT-Interaction Dataset, http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html
[16]. MSR Action Dataset, http://research.microsoft.com/en-us/downloads/fbf24c35-a93e-4d22-a5fe-bc08f1c3315e/

