Survey on expedition of Human Activity detection and recognition algorithm.

MS. Kanchan Gaikwad
Department of Computer,
MGM College of Engineering,
Kamothe, Panvel.

Prof M r.Vaibhav Narwade
Department of Information Technology,
PVPPSCOE, Sion, Mumbai

Abstract:

Visual surveillance is an active research topic in image processing. Transit systems are actively seeking new or improved ways to use technology to prevent and respond to suspicious activities accidents, crime, suspicious activities, terrorism, and vandalism. Human behavior-recognition algorithms can be used for prevention of incidents or reactively for investigation after the fact. This paper describes the survey on journey of human activity detection and recognition algorithm, it shows current state-of-the-art image-processing methods for automatic-behavior-recognition techniques. The main goal of this survey is to provide detail information of researches are done till date in this area. Human behavior-recognition methods for transit surveillance. Recognition methods include single person (e.g., loitering), multiple person interactions (e.g., fighting and personal attacks), person–vehicle interactions (e.g., vehicle vandalism), and person–facility/location interactions (e.g., object left behind and trespassing). This paper is also include the various application of human activity recognition in real market.

Index Terms: Human activity recognition, HMM techniques, Thresholding,

Section 1

Introduction

Recognizing human activities from video is one of the most promising applications of computer vision. In recent years, this problem has caught the attention of researchers from industry, academia, security agencies, consumer agencies and the general populace too. Automatic Human Activity Recognition (HAR) has received great attention by researchers involved in human – computer interaction, due to the continuous need for smarter and more user - friendly interfaces. The analysis of human body movements can be applied in a variety of application domains, such as video surveillance, video retrieval, human computer interaction systems, and medical diagnoses. In some cases, the results of such analysis can be applied to identify suspicious action of people and other unusual events automatically from the videos, without any human intrusion.

The terms ‘Action’ and ‘Activity’ are frequently used interchangeably in the vision literature. In the ensuing discussion, by ‘Actions’ we refer to simple motion patterns usually executed by a single person and typically lasting for short durations of time, on the order of tens of seconds. Examples of actions include bending, walking, swimming etc. On the other hand, by ‘Activities’ we refer to the complex sequence of actions performed by several humans who could be interacting with each other in a constrained manner. They are typically characterized by much longer temporal durations, e.g. two persons shaking hands, a football team scoring a goal or a coordinated bank attack by multiple robbers. This is not a hard boundary and there is a significant ‘gray-area’ between these two extremes.

This paper is arranged as section 1 gives introduction of paper. The overview

about technique used for recognition of human motion is including in section 2. The various application of Human Activity Recognition (HAR) is include in section 3.The section 4 include the conclusion of all technique used in Human Activity Recognition (HAR).

Section 2

Overview of Technique used for recognition

The terms 'Action' and 'Activity' are frequently used interchangeably in the vision literature. A generic action or activity recognition system can be viewed as proceeding from a sequence of images to a higher level interpretation in a series of steps. The major steps involved are the following:

1) Input video or sequence of images
2) Extraction of concise low-level features
3) Mid-level action descriptions from low-level futures
4) High-level semantic interpretations from primitive actions.

When we discuss about 'Actions' we usually refer to simple motion patterns which executed by a single person and typically for short durations of time, on the order of tens of seconds. Examples of actions include bending, walking, swimming etc.

On the other hand, for 'Activities' we usually refer to the complex sequence of actions performed by several humans who could be interacting with each other in a constrained manner. They are typically characterized by much longer temporal durations, e.g. two persons shaking hands or two persons fighting with each other. There are so many approaches for detecting activity and action, depending on their categories the relative approaches get select. For e.g. Real life activity recognition systems typically follow a hierarchical approach. At the lower levels are modules such as background foreground segmentation, tracking and object detection. At the mid-level are action-recognition modules. At the high-level are the reasoning engines which encode the activity semantics based on the lower level action-primitives. Thus it is necessary to gain an understanding of both these problem domains to enable real-life deployment of systems [1]. A quick preview of the various approaches that fall under each of these categories is shown in figure 1.

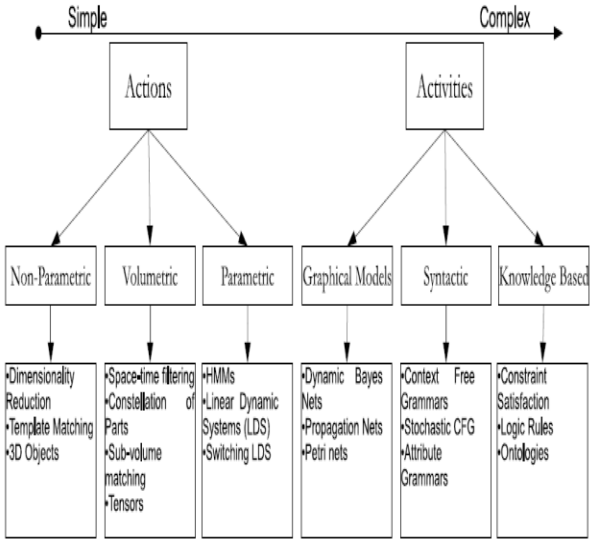

Fig1: Overview of approaches for activity and action recognition.

Methods for recognizing actions:

i) Non Parametric Approach

2D-templates: One of the earliest attempts at action recognition without relying on 3-D structure estimation was proposed by Polana and Nelson [2]. First, they perform motion-detection and tracking of humans in the scene. After tracking, a 'cropped' sequence containing the human is constructed. Scale changes are compensated for by normalizing

the size of the human. A periodicity index is computed for the given action and the algorithm proceeds to recognize the action if it is found to be sufficiently periodic. To perform recognition, the periodic sequence is segmented into individual cycles using the periodicity estimate and combined to get an average-cycle. The average-cycle is divided into a few temporal segments and flow based features are computed for each spatial location in each segment. The flow-features in each segment are averaged into a single frame. The average flow frames within an activity-cycle form the templates for each action class. Bobick and Davis [3] proposed 'temporal templates' as models for actions. In their approach, the first step involved is background subtraction, followed by an aggregation of a sequence of background subtracted blobs into a single static image. They propose two methods of aggregation – the first method gives equal weight to all images in the sequence, which gives rise to a representation called the 'Motion Energy Image' (MEI). The second method gives decaying weights to the images in the sequence with higher weight given to new frames and low weight to older frames. This leads to a representation called the 'Motion History Image' (MHI).

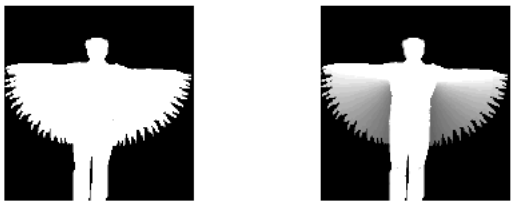

Fig. 2. Temporal templates similar to [3]. Left: Motion Energy Image of a sequence of a person raising both hands, Right: Motion History Image of the same action.

The MEI and MHI together comprise a template for a given action. From the templates, translation, rotation and scale invariant Hu-moments [4] are extracted which are then used for recognition. It was shown in [3] that MEI and MHI have sufficient discriminating ability for several simple action classes such as 'sitting down', 'bending', 'crouching' and other aerobic postures. However, it was noted in [5] that MEI and MHI lose discriminative power for complex activities due to over-writing of the motion history and hence are unreliable for matching.

3D Object models: Successful application of models and algorithms to object recognition problems led researchers in action recognition to propose alternate representations of actions as spatio temporal objects. Syeda-Mahmood et al. proposed a representation of actions as generalized cylinders in the joint (x, y, t) space [6]. Yilmaz and Shah [7] represent actions as 3-D objects induced by stacking together tracked 2- D object contours. A sequence of 2-D contours in (x, y) space can be treated as an object in the joint (x, y, t) space. This representation encodes both the shape and motion characteristics of the human. From the (x, y, t) representation, concise descriptors of the object's surface are extracted corresponding to geometric features such as peaks, pits, valleys and ridges. Since this approach is based on stacking together a sequence of silhouettes, accurate correspondence between points of successive silhouettes in the sequences needs to be established. Quasi view invariance for this representation was shown theoretically by assuming an affine camera model. Similar to this approach, [8] proposed using background subtracted blobs instead of contours, which are then stacked together to create an (x, y, t) binary space-time volume Since this approach uses background subtracted blobs, the problem of establishing correspondence between points on contours in the sequence does not exist. From this space time volume, 3-D shape descriptors are extracted by solving a Poisson equation [8]. Since these approaches

require careful segmentation of background and the foreground, they are
limited in applicability to fixed camera settings.

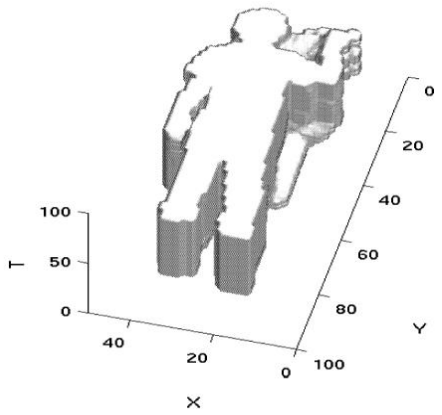

Fig. 3. 3D space-time object, similar to [7], obtained by stacking together binary background subtracted images of a person waving his hand.

II. Volumetric Approaches

1) *Spatio-temporal Filtering:* These approaches are based on filtering a video volume using a large filter bank. The responses of the filter bank are further processed to derive action specific features. These approaches are inspired by the success of filter-based methods on other still image recognition tasks such as texture segmentation [9]. Further, spatiotemporal filter structures such as oriented Gaussian kernels and their derivatives [10] and oriented Gabor filter banks [11] have been hypothesized to describe the major spatiotemporal properties of cells in the visual cortex. Chomat et al. [12] model a segment of video as a (x, y, t) spatiotemporal volume and compute local appearance models at each pixel using a Gabor filter bank at various orientation and spatial scales and a single temporal scale. A given action is recognized using a spatial average of the probabilities of individual pixels in a frame. Since actions are analyzed at a single temporal scale, this method is not applicable to variations in execution rate.

2) *Part-Based Approaches:* Several approaches have been proposed that consider a video volume as a collection of local parts, where each part consists of some distinctive motion pattern. Laptev and Lindeberg [13] proposed a spatiotemporal generalization of the well-known Harris interest point detector, which is widely used in object recognition applications and applied it to modeling and recognizing actions in space-time. This method is based on the 3D generalization of scale space representations. A given video is convolved with a 3D Gaussian kernel at various spatial and temporal scales. Then, spatiotemporal gradients are computed at each level of the scale-space representation. These are then combined within a neighborhood of each point to yield stable estimates of the spatiotemporal second moment matrix.

III. Parametric Methods

1) *Hidden Markov Models:* One of the most popular state space models is the Hidden Markov Model. HMMs are efficient for modeling time-sequence data and are useful both for their generative and discriminative capabilities. HMMs are well-suited for tasks that require recursive probabilistic estimates [63] or when accurate start and end times for action units are unknown. However, their utility is restricted due to the

simplifying assumptions that the model is based on. Most significantly the assumption of Markovian dynamics and the time invariant nature of the model restricts the applicability of HMMs to relatively simple and *stationary* temporal patterns.

2) *Linear Dynamical Systems:* Linear dynamical systems are a more general form of HMMs where the state-space is not constrained to be a finite set of symbols but can take on continuous values in R^k where k is the dimensionality of the state-space. The simplest form of LDS is the first order time-invariant Gauss-Markov processes which is described by equations (1) and (2)

$$x(t) = Ax(t-1) + w(t),\ w \sim N(0,Q) \quad (1)$$

$$y(t) = Cx(t) + v(t),\ v \sim N(0,R) \quad (2)$$

where $x \in R^d$ is the d dimensional state vector and $y \in R^n$ is the n-dimensional observation vector with $d << n$. w and v are the process and observation noise respectively which are Gaussian distributed with zero-means and covariance matrices Q and R respectively.

Section III

Application of Human Activity Recognition

In this section, we focusing on a few application areas that will highly the potential impact of vision-based activity recognition systems.

1) Behavioral Biometrics: It involves study of approaches and algorithms for uniquely recognizing humans based on physical or behavioral cues. Traditional approaches are based on fingerprint, face or iris and can be classified as Physiological Biometrics i.e. they rely on physical attributes for recognition. These methods require cooperation from the subject for collection of the biometric. Recently, 'Behavioral Biometrics' have been gaining popularity, where the premise is that behavior is as useful a cue to recognize humans as their physical attributes. The advantage of this approach is that subject-cooperation is not necessary and it can proceed without interrupting or interfering with the subject's activity. Since observing behavior implies longer-term observation of the subject, approaches for action-recognition extend naturally to this task. Currently, the most promising example of behavioral biometric is human gait [15].

2) Content Based Video Analysis: Video has become a part of our everyday life. With video sharing websites experiencing relentless growth, it has become necessary to develop efficient indexing and storage schemes to improve user experience. This requires learning of patterns from raw video and summarizing a video based on its content. Content-based video summarization has been gaining renewed interest with corresponding advances in content-based image retrieval (CBIR) [16]. Summarization and retrieval of consumer content such as sports videos is one of the most commercially viable applications of this technology [17].

3) Security and Surveillance: Security and surveillance systems have traditionally relied on a network of video cameras monitored by a human operator who needs to be aware of the activity in the camera's field of view. With recent growth in the number of cameras and deployments, the efficiency and accuracy of human operators has been stretched. Hence, security agencies are seeking vision-based solutions to these tasks which can replace or assist a human operator. Automatic recognition of anomalies in a camera's field of view is one such problem that has attracted attention from vision researchers [18]. A related application involves

searching for an activity of interest in a large database by learning patterns of activity from long videos [19], [20].

4) Interactive Applications and Environments: Understanding the interaction between a computer and a human remains one of the enduring challenges in designing human-computer interfaces. Visual cues are the most important mode of nonverbal communication. Effective utilization of this mode such as gestures and activity holds the promise of helping in creating computers that can better interact with humans. Similarly, interactive environments such as smart rooms [21] that can react to a user's gestures can benefit from vision based methods. However, such technologies are still not mature enough to stand the 'Turing test' and thus continue to attract research interest.

5) Animation and Synthesis: The gaming and animation industry rely on synthesizing realistic humans and human motion. Motion synthesis finds wide use in the gaming industry where the requirement is to produce a large variety of motions with some compromise on the quality. The movie industry on the other hand has traditionally relied more on human animators to provide high-quality animation. However, this trend is fast changing [22]. With improvements in algorithms and hardware, much more realistic motion-synthesis is now possible. A related application is learning in simulated environments. Examples of this include training of military soldiers, fire-fighters and other rescue personnel in hazardous situations with simulated subjects.

Section 3

Conclusion

The technique used for Human activity recognition are vary as per the process or methods used for image filtration. The Basic steps for Human activity recognition from video is first the video is get convert into continues frame and then the difference between two frame is consider as the movement by object. There are the different methods for filtering the image and finding the concise low-level features then mid level action descriptions from low-level features. The Hierarchical syntactic approach is useful for activities with deep hierarchical structure and repetitive (cyclic) structure. Context free grammar (CFG) is good for structured activities it can incorporate uncertainty in observations and natural contextual prior for recognizing errors. Hierarchical statistical approach is used when Low-level action detectors are noisy; Structure of activity is sequential and integrating dynamics. In case with HMM it gives result with minimum number of frames as compare with other within less time.

REFERENCES

1. J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, 1999.
2. R. Polana and R. C. Nelson, "Detection and recognition of periodic, nonrigid motion," *International Journal of Computer Vision*, vol. 23,no. 3, pp. 261–282, 1997.
3 A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

4 M.-K. Hu, "Visual pattern recognition by moment invariants," *IEEE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.
5 A. F. Bobick, "Movement, activity, and action: The role of knowledge in the perception of motion," p*hilosophical Transactions of the Royal Society of London B*, vol. 352, pp. 1257–1265, 1997.
6 A. F. Bobick, "Movement, activity, and action: The role of knowledge in the perception of motion," *Philosophical Transactions of the Royal Society of London B*, vol. 352, pp. 1257–1265, 1997.
7 T. F. Syeda-Mahmood, M. Vasilescu, and S. Sethi, "Recognizing action events from multiple viewpoints," *IEEE Workshop on Detection and Recognition of Events in Video*, pp. 64–72, 2001.
8 Y. Yacoob and M. J. Black, "Parameterized modeling and recognition of activities," *Computer Vision and Image Understanding*, vol. 73, no. 2, pp. 232–247, 1999.
9 J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanism," *Journal of Optical Society of America-A*, vol. 7, no. 5, pp. 923–932, May 1990.
10 R. A. Young, R. M. Lesperance, and W. W. Meyer, "The Gaussian derivative model for spatial-temporal vision: I. cortical model," *Spatial Vision*, vol. 14, no. 3–4, pp. 261–319, 2001.
11 H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," *Proceedings of IEEE InternationalConference on Computer Vision*, 2007.
12 O. Chomat and J. L. Crowley, "Probabilistic recognition of activity using local appearance," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 02, pp. 104–109, 1999.
13 I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.
14 J. Schlenzig, E. Hunter, and R. Jain, "Recursive identification of gestureinputs using hidden markov models," *Proceedings of the Second IEEEWorkshop on Applications of Computer Vision*, pp. 187–194, 1994.
15 [10] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The HumanID gait challenge problem: Data sets, performance, and analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 162–177, 2005.
16 Y. Rui, T. S. Huang, and S. F. Chang, "Image retrieval: current techniques, promising directions and open issues," *Journal of Visual Communication and Image Representation*, vol. 10, no. 4, pp. 39–62, 1999.
17 S. F. Chang, "The holy grail of content-based media analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 6–10, 2002.
18 H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 819–826, 2004.
19 C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.
20 W. Hu, D. Xie, T. Tan, and S. Maybank, "Learning activity patterns using fuzzy self-organizing neural network," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 34, no. 3, pp. 1618–1626, 2004.
21 A. Pentland, "Smart rooms, smart clothes," *International Conference on Pattern Recognition*, vol. 2, pp. 949–953, 1998.
22 D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan, "Computational studies of human motion: part 1, tracking and motion synthesis," *Foundations and Trends in Computer Graphics and Vision*, vol. 1, no. 2-3, pp. 77–254, 2005.