Machine Recognition of Human Activities: A survey

Pavan Turaga, *Student Member*, Rama Chellappa, *Fellow, IEEE*, V. S. Subrahmanian, and Octavian Udrea, *Student Member*

***Abstract*— The past decade has witnessed a rapid proliferation of video cameras in all walks of life and has resulted in a tremendous explosion of video content. Several applications such as content-based video annotation and retrieval, highlight extraction and video summarization require recognition of the activities occurring in the video. The analysis of human activities in videos is an area with increasingly important consequences from security and surveillance to entertainment and personal archiving. Several challenges at various levels of processing - robustness against errors in low-level processing, view and rate-invariant representations at mid-level processing and semantic representation of human activities at higher-level processing – make this problem hard to solve. In this review paper, we present a comprehensive survey of efforts in the past couple of decades to address the problems of representation, recognition and learning of human activities from video and related applications. We discuss the problem at two major levels of complexity – a) 'actions' and b) 'activities'. 'Actions' are characterized by simple motion-patterns typically executed by a single human. 'Activities' are more complex and involve co-ordinated actions among a small number of humans. We shall discuss several approaches and classify them according to their ability to handle varying degrees of complexity as interpreted above. We begin with a discussion of approaches to model the simplest of action classes known as atomic or primitive actions which do not require sophisticated dynamical modeling. Then, methods to model actions with more complex dynamics are discussed. The discussion then leads naturally to methods for higher-level representation of complex activities.**

I. 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. One of the earliest investigations into the nature of human motion was conducted by the contemporary photographers Etienne Jules Marey and Eadweard Muybridge in the 1850s who photographed moving subjects and revealed several interesting and artistic aspects involved in human and animal locomotion. The classic Moving Light Display (MLD) experiment of Johansson [1] provided a great impetus to the study and analysis of human motion perception in the field of neuroscience. This then paved the way for mathematical modeling of human action and automatic recognition, which naturally fall into the purview of computer vision and pattern recognition.

To state the problem in simple terms, given a sequence of images with one or more persons performing an activity, can a system be designed that can automatically recognize what activity is being or was performed ? As simple as the question seems, the solution has been that much harder to find. In this survey paper we review the major approaches that have been pursued over the last 20 years to address this problem.

Several related survey papers have appeared over the years. Most notable among them are the following. Aggarwal and Cai [2] discuss three important sub-problems that together form a complete action recognition system – extraction of human body structure from images, tracking across frames and action-recognition. Cedras and Shah [3] present a survey on motion-based approaches to recognition as opposed to structure-based approaches. They argue that motion is a more important cue for action recognition than the structure of the human body. Gavrila [4] presented a survey focused mainly on tracking of hands and humans via 2D or 3D models and a discussion of action recognition techniques. More recently, Moeslund et al [5] presented a survey of problems and approaches in human motion capture including human model initialization, tracking, pose estimation and activity recognition. Since the mid-90s, interest has shifted more toward recognizing actions from tracked motion or structure features and on recognizing complex activities in real-world settings. Considerable effort has been spent on addressing these problems in the last several years. Hence, this survey will focus exclusively on approaches for recognition of action and activities from video and not on the lower-level modules of detection and tracking which is discussed at length in earlier surveys [2], [3], [4], [5], [6].

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 (e.g. figure 1). 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 co-ordinated bank attack by multiple robbers (figure 2). This is not a hard boundary and there is a significant 'gray-area' between these two extremes. For example, the gestures of a music conductor conducting an orchestra or the constrained

 Pavan Turaga, Rama Chellappa, V. S. Subrahmanian and Octavian Udrea are with the Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742. Email: $\{pturaga, rama, vs, udrea\}$@$umiacs.umd.edu$. This work was funded in part by the U.S. Government VACE program.

dynamics of a group of humans (figure 3) is neither as simple as an 'action' nor as complex as an 'activity' according to the above interpretation. However, this simple categorization provides a starting-point to organize the numerous approaches that have been proposed to solve the problem. A quick preview of the various approaches that fall under each of these categories is shown in figure 4. The action and activity recognition approaches are complementary to each other. Real-life activity recognition systems typically follow a hierarchical approach. At the lower levels are standard vision 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/structure 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.

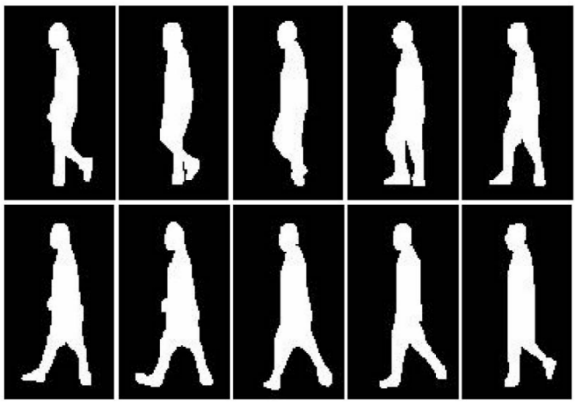

Fig. 1. Near-field video: Example of Walking action. Figure taken from [7].

Fig. 3. Far-field video: Modeling dynamics of groups of humans as a deforming shape. Figure taken from [9].

The rest of the paper is organized as follows. First, we discuss a few motivating application domains in section II. Section III provides an overview of methods for extraction of low-level image features. In section IV we discuss approaches for recognizing 'actions'. Then, in section V we discuss methods to represent and recognize higher-level 'activities'. In section VI, we discuss some open research issues for action and activity recognition and provide concluding remarks.

II. Applications

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

1) Behavioral Biometrics: Biometrics 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 [10].

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) [11]. Summarization and retrieval of consumer content such as sports videos is one of the most commercially viable applications of this technology [12].

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 (c. f. [13], [9]). A related application involves searching for an activity of interest in a large database by learning patterns of activity from long videos [14], [15].

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 non-verbal 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 [16] 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

Fig. 2. Medium-field video: Example video sequence of a simulated bank attack (courtesy [8]). (a) Person enters the bank, (b) Robber is identified to be an outsider. Robber is entering the bank safe, (c) A customer escapes, (d) Robber makes an exit.

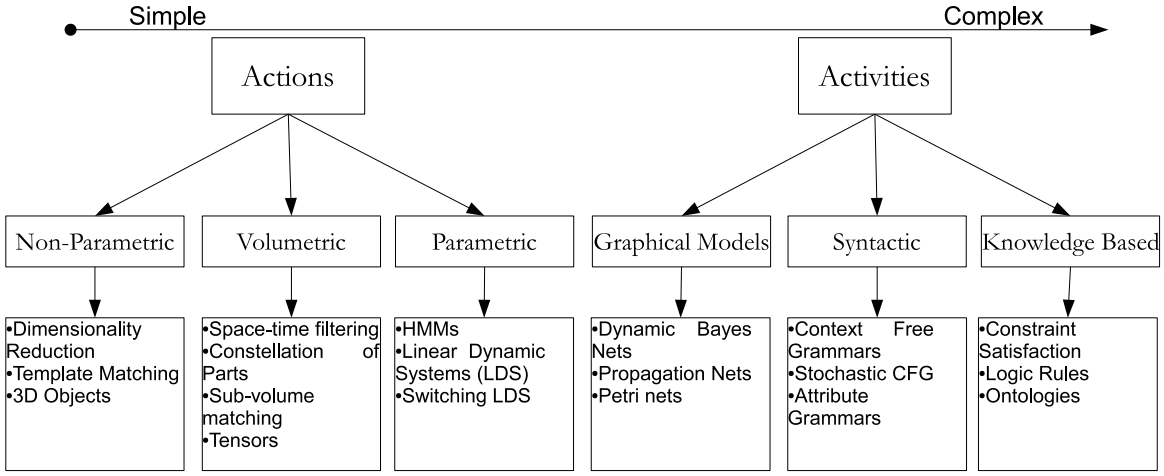

Fig. 4. Overview of approaches for action and activity recognition.

human animators to provide high-quality animation. However, this trend is fast changing [17]. 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.

III. General Overview

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 features
4) High-level semantic interpretations from primitive actions

In this section, we will briefly discuss some relevant aspects of item 2, i.e. low-level feature extraction. Items 3 and 4 in the list will form the subject of discussion of sections IV and V respectively.

Videos consist of massive amounts of raw information in the form of spatio-temporal pixel intensity variations. But most of this information is not directly relevant to the task of understanding and identifying the activity occurring in the video. A classic experiment by Johansson [1] demonstrated that humans can perceive gait patterns from point light sources placed at a few limb joints with no additional information. Extraneous factors such as the color of the clothes, illumination conditions, background clutter do not aid in the recognition task. We briefly describe a few popular low-level features and refer the readers to other sources for a more in-depth treatment as we progress.

A. Optical flow

Optical flow is defined as the apparent motion of individual pixels on the image plane. Optical flow often serves as a good approximation of the true physical motion projected onto the image plane. Most methods to compute optical flow assume that the color/intensity of a pixel is invariant under the displacement from one video frame to the next. We refer the reader to [18] for a comprehensive survey and comparison of optical flow computation techniques. Optical flow provides a concise description of both the regions of the image undergoing motion and the velocity of motion. In practice, computation of optical flow is susceptible to noise and illumination changes. Applications include [19] which used optical flow to detect and track vehicles in an automated traffic surveillance application.

B. Point trajectories

Trajectories of moving objects have popularly been used as features to infer the activity of the object (see figure 5). The image-plane trajectory itself is not very useful as it is sensitive to translations, rotations and scale changes. Alternative representations such as trajectory velocities, trajectory speeds, spatio-temporal curvature, relative-motion etc have

been proposed that are invariant to some of these variabilities. A good survey of these approaches can be found in [3]. Extracting unambiguous point trajectories from video is complicated by several factors such as occlusions, noise and background clutter. Accurate tracking algorithms need to be employed for obtaining motion trajectories [6].

Fig. 5. Trajectories of a passenger and luggage-cart. The wide difference in the trajectories is indicative of the difference in activities. Figure taken from [20].

C. Background subtracted blobs and Shape

Background subtraction is a popular method to isolate the moving parts of a scene by segmenting it into background and foreground. As an example, from the sequence of background subtracted images shown in figure 1, the human's walking action can be easily perceived. Several approaches to background modeling exist. One popular approach is to learn a statistical distribution of pixel intensities that correspond to the background as in [21]. By adapting the background model according to new data, the method can also be applied to scenarios with changing background [21]. Shape of the human silhouette plays a very important role in recognizing human actions (see figure 6). Several methods based on global, boundary and skeletal descriptors have been proposed to quantify shape. Global methods such as moments [22] consider the entire shape region to compute the shape-descriptor. Boundary methods on the other hand consider only the shape contour as the defining characteristic of the shape. Such methods include chain codes [23] and landmark-based shape descriptors [24]. Skeletal methods represent a complex shape as a set of 1D skeletal curves, for example, the medial axis transform [25]. Applications include shape-based dynamic modeling of the human silhouette as in [26] to perform gait recognition.

D. Filter Responses

There are several other features which can be broadly classified as based on spatio-temporal filter responses. In their work, Zhong et al [13] process a video sequence using a spatial Gaussian and a derivative of Gaussian on the temporal axis. Due to the derivative operation on the temporal axis, the filter shows high responses at regions of motion. This response was then thresholded to yield a binary motion mask followed by aggregation into spatial histogram bins. Such a feature encodes motion and its corresponding spatial information compactly and is useful for far-field and medium-field surveillance videos. The notion of scale-space filtering has also been extended to videos by several researchers. Laptev et al [27], [28] propose a generalization of the Harris corner detector to videos using a set of spatio-temporal Gaussian derivative filters. Similarly, Dollar et al [29] extract distinctive periodic motion-based landmarks in a given video using a Gaussian kernel in space and a Gabor function in time. Since these approaches are based on simple convolution operations, they are fast and easy to implement. They are quite useful in scenarios with low resolution or poor quality video where it is difficult to extract other features such as optical flow or silhouettes.

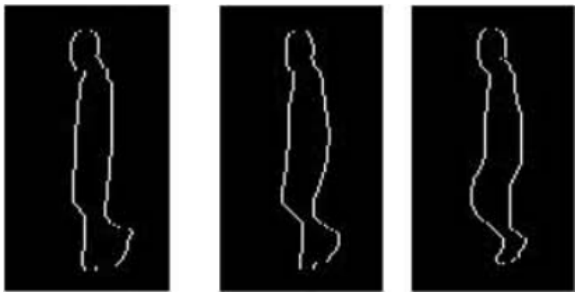

Fig. 6. Silhouettes extracted from the walking sequence shown in figure 1. Silhouettes encode sufficient information to recognize actions. Figure taken from [7].

IV. Modeling and Recognizing Actions

Approaches for human action recognition fall into one of the two following categories – a) Methods that rely on human body models, b) Methods that do not rely on human body models. Methods that fall in the first category rely on segmentation of the body into individual parts and extract features such as joint-angles or joint-trajectories. Segmentation of the human body is a computationally intensive task and extraction of joint trajectories requires good tracking algorithms. These approaches were popular in the early 90s and an excellent survey can be found in [2]. More recently, the focus has shifted toward approaches which do not assume a body model. These methods rely on motion information extracted directly from the images. Motion-based approaches for modeling actions fall into three major classes – non-parametric, volumetric and parametric time-series approaches. Non-parametric approaches typically extract a set of features from each frame of the video. The features are then matched to a stored template. Volumetric approaches on the other hand do not extract features on a frame-by-frame basis. Instead, they consider a video as a 3D volume of pixel intensities and extend standard image features such as scale-space extrema, spatial filter responses etc to the 3D case. Parametric time-series approaches specifically impose a model on the temporal dynamics of the motion. The particular parameters for a class of actions is then estimated from training data. Examples of parametric approaches include Hidden Markov Models (HMMs), Linear Dynamical Systems (LDSs) etc. We will first discuss the non-parametric methods, then the volumetric approaches and finally the parametric time-series methods.

A. Non-Parametric Approaches for Action Recognition

1) 2D-templates: One of the earliest attempts at action-recognition without relying on 3-D structure estimation was proposed by Polana and Nelson [30], [31]. 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. Other related approaches for representation and recognition of quasi-cyclic actions have been proposed in [32], [33]. Since these methods are based on periodic motion, they are best suited to quasi-periodic actions such as walking, running, swimming etc.

Bobick and Davis [34], [35] 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) (for example, see figure 7). The MEI and MHI together comprise a template for a given action. From the templates, translation, rotation and scale invariant Hu-moments [22] are extracted which are then used for recognition. It was shown in [35] 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 [36] that MEI and MHI lose discriminative power for complex activities due to over-writing of the motion history and hence are unreliable for matching.

2) 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 [37]. Yilmaz and Shah [38] 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

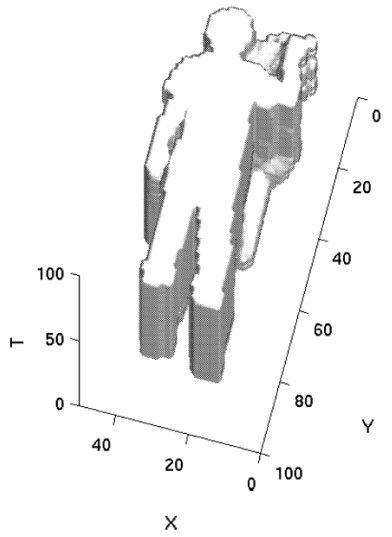

Fig. 8. 3D space-time object, similar to [39], obtained by stacking together binary background subtracted images of a person waving his hand.

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, [39], [40] proposed using background subtracted blobs instead of contours, which are then stacked together to create an (x, y, t) binary space-time volume (for example, see figure 8). 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 [39], [40]. Since these approaches require careful segmentation of background and the foreground, they are limited in applicability to fixed camera settings.

3) Manifold Learning Methods: Most approaches in action recognition involve dealing with data in very high-dimensional spaces. Hence, these approaches often suffer from the 'curse of dimensionality'. The feature-space becomes sparser in an exponential fashion with the dimension, thus requiring a larger number of samples to build efficient class-conditional models. Learning the manifold on which the data resides enables us to determine the inherent dimensionality of the data as opposed to the raw dimensionality. The inherent dimensionality contains fewer degrees of freedom and allows efficient models to be designed in the lower-dimensional space. The simplest way to reduce dimensionality is via Principal Component Analysis (PCA) which assumes that the data lies on a linear subspace. Except in very special cases, data does not lie on a linear subspace, thus requiring methods that can learn the intrinsic geometry of the manifold from a large number of samples. Nonlinear dimensionality reduction techniques allow for representation of data points based on their proximity to each other on nonlinear manifolds. Several methods for dimensionality reduction such as PCA, locally linear embedding (LLE) [41], Laplacian eigenmap [42], and Isomap [43] have been applied to reduce the high-dimensionality of video data in action-recognition tasks (c. f. [44], [45], [46]). Specific recognition algorithms such as template matching, dynamical modeling etc

Fig. 7. Temporal templates similar to [34]. Left: Motion Energy Image of a sequence of a person raising both hands, Right: Motion History Image of the same action.

can be performed more efficiently once the dimensionality of the data has been reduced.

B. Volumetric Approaches

The approaches discussed above relied on extracting features such as shape and optical-flow from each frame individually, hence are limited to conditions where these features can be reliably extracted. A different approach is to analyze chunks of video in 3D directly as opposed to sequentially analyzing frames. This is possible in large part due to advances in processing power and memory capacities of modern computers. Further, these methods do not require sophisticated segmentation and tracking algorithms, instead relying on extracting volumetric features. In this section, we review some of the approaches that directly extend the 2D approaches to 3D video volumes.

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 [47]. Further, spatio-temporal filter structures such as oriented Gaussian kernels and their derivatives [48] and oriented Gabor filter banks [49] have been hypothesized to describe the major spatio-temporal properties of cells in the visual cortex. Chomat et al. [50] model a segment of video as a (x, y, t) spatio-temporal 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. As an extension to this approach, local histograms of normalized space-time gradients at several temporal scales are extracted by Zelnik-Manor and Irani [51]. The sum of the chi-square metric between histograms is used to match an input video with a stored exemplar. Filtering with the Gaussian kernel in space and the derivative of the Gaussian on the temporal axis followed by thresholding of the responses and accumulation into spatial-histograms was found to be a simple yet effective feature for actions in a far field settings [13].

Filtering approaches are fast and easy to implement due to efficient algorithms for convolution. In most applications the appropriate bandwidth of the filters is not known a priori, thus a large filter bank at several spatial and temporal scales is required for effectively capturing the action dynamics. Moreover the response generated by each filter has the same dimensions as the input volume, hence using large filter banks at several spatial and temporal scales is prohibitive.

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 [27], [28] proposed a spatio-temporal 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, spatio-temporal 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 spatio-temporal second-moment matrix. Local features are then derived from these smoothed estimates of gradient moment matrices. In a similar approach, Dollar et al. [29] model a video sequence by the distribution of space-time (ST) feature prototypes. The feature prototypes are obtained by k-means clustering of a large set of features – space-time gradients – extracted at ST interest points from the training data. Neibles et al. [52] use a similar approach where they use a bag-of-words model to represent actions. The bag-of-words model is learnt by extracting spatio-temporal interest points and clustering of the features. These interest points can be used in conjunction with machine learning approaches such as SVMs [53] and graphical models [52]. Since the interest-points are local in nature, longer term temporal correlations are ignored in these approaches. To address this issue, a method based on correlograms of prototype labels was presented in [54]. In a slightly different approach Nowozin et al [55] consider a video as a sequence of sets - where each set consists of the parts found in a small temporally sliding window. These approaches do not directly model the global geometry of local parts instead considering them as a bag-of-features. Different actions may be composed of similar space-time parts but may differ in their geometric relationships. Integrating global geometry into the part-based video representation was investigated by Boiman et al [56] and Wong et al [57]. This

approach may be termed as a constellation-of-parts as opposed to the simpler bag-of-parts model. Computational complexity can be large for constellation models with a large number of parts which is typically the case for human actions. Song et al [58] addressed this issue by approximating the connections in the constellation via triangulation. Niebles et al [59] proposed a hierarchical model where the higher level is a constellation of parts much smaller than the actual number of features. Each of the parts in the constellation consists of a bag-of-features at the lower level. This approach combines the advantages of both the bag-of-features and the constellation model and preserves computational efficiency at the same time.

In most of these approaches the detection of the parts is usually based on linear operations such as filtering and spatio-temporal gradients, hence the descriptors are sensitive to changes in appearance, noise, occlusions etc. It has also been noted that interest points are extremely sparse in smooth human actions and certain types of actions do not give rise to distinctive features [52], [29]. However, due to their local nature they are more robust to non-stationary backgrounds.

3) Sub-volume matching: As opposed to part-based approaches, researchers have also investigated matching of videos by matching sub-volumes between a video and a template. Shechtman et al [60] present an approach derived from space-time motion based correlation to match actions with a template. The main difference of this approach from the part-based approaches is that it does not extract action descriptors from extrema in scale-space, rather it looks for similarity between local space-time patches based on how similar the motion is in the two patches. However, computing this correlation throughout a given video volume can be computationally intensive. Inspired by the success of Haar-type features or 'box-features' in object detection [61], Ke et al [62] extended this framework to 3D. In their approach, they define 3D Haar-type features which are essentially outputs of 3D filter banks with $+1$'s and -1's as the filter co-efficients. The filters themselves are very coarse and simple in nature but when several of these filters are convolved with a given video at various spatial and temporal scales and used in conjunction with boosting approaches, very robust performance is obtained. They also propose an efficient means of extracting the large number box-features using a generalization of the integral-image to integral-video. In another approach, Ke et al [63] consider a video volume as a collection of sub-volumes of arbitrary shape, where each subvolume is a spatially coherent region. The subvolumes are obtained by clustering the pixels based on appearance and spatial proximity. A given video is over-segmented into many subvolumes or 'supervoxels'. An action template is matched by searching among the over-segmented volumetric regions and finding the minimal set of regions that maximize overlap between their union and the template.

Sub-volume matching approaches such as these are susceptible to changing backgrounds but are more robust to noise and occlusions. Another advantage is that these approaches can be extended to features such as optical flow as in [62] to achieve robustness to changes in appearance.

4) Tensor based approaches: Tensors are generalizations of matrices to multiple dimensions. A 3D space-time volume can naturally be considered as a tensor with three independent dimensions. Vasilescu [64] proposed the modeling of human action, human identity and joint angle trajectories by considering them as independent dimensions of a tensor. By decomposing the overall data tensor into dominant modes (as a generalization of principal component analysis), one can extract signatures corresponding to both the action and the identity of the person performing the action. Recently, Kim et al [65] extended canonical correlation analysis to tensors to match videos directly to templates. In their approach, the dimensions of the tensor were simply the space-time dimensions corresponding to (x, y, t). Similarly, Wolf et al [66] extended low-rank SVM techniques to the space of tensors for action recognition.

Tensor-based approaches offer a direct method for holistic matching of videos without recourse to mid-level representations such as the previous ones. Moreover, they can incorporate other types of features such as optical flow, space-time filter responses etc into the same framework by simply adding more independent dimensions to the tensor.

C. Parametric Methods

The previous section focused on representations and models for simple actions – known as atomic or primitive actions. These approaches can model short-term actions but are not well suited for temporally extended actions. The parametric approaches that we will describe in this section are better suited for these actions. Parametric methods such as HMMs and LDSs are well suited to model more complex actions where the underlying process is characterized by complex temporal dynamics. In such cases, simple template matching approaches would either require too many templates or would not capture the dynamics of the action at all. Examples of such complex actions include the steps in a ballet dancing video, a juggler juggling a ball and a music conductor conducting an orchestra using complex hand gestures.

The most popular method used for modeling complex temporal dynamics are the so called state-space approaches. State-space approaches model the temporal evolution of features as a trajectory in some configuration space, where each point on the trajectory corresponds to a particular 'configuration' or 'state' – for instance, a particular pose or stance of the actor.

1) Hidden Markov Models: One of the most popular state-space models is the Hidden Markov Model. In the discrete HMM formalism, the state space is considered to be a finite set of discrete points. The temporal evolution is modeled as a sequence of probabilistic jumps from one discrete state to the other (figure 9). HMMs first found wide applicability in speech recognition applications in the early 80s. An excellent source for a detailed explanation of HMMs and its associated three problems – inference, decoding and learning – can be found in [67]. Beginning in the early 90's, HMMs began to find wide applicability in computer vision systems. One of the earliest approaches to recognize human actions via HMMs was proposed by Yamato et al. [68] where they recognized tennis

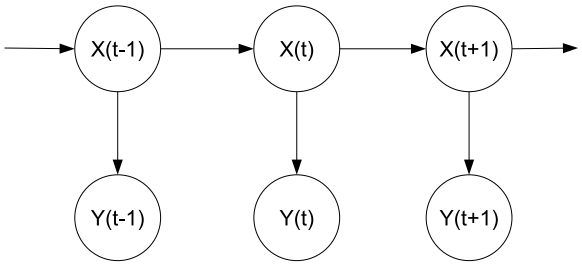

Fig. 9. Graphical Illustration of a Hidden Markov Model.

shots such as backhand stroke, backhand volley, forehand stroke, forehand volley, smash etc by modeling a sequence of background subtracted images as outputs of class-specific HMMs. Several successful gesture recognition systems such as in [69], [70], [71] make extensive use of HMMs by modeling a sequence of tracked features such as hand blobs as HMM outputs.

HMMs have also found applicability in modeling the temporal evolution of human gait patterns both for action recognition and biometrics (cf. Kale et al. [72], Liu and Sarkar [73]). All these approaches are based on the assumption that the feature sequence being modeled is a result of a single person performing an action. Hence, they are not effective in applications where there are multiple agents performing an action or interacting with each other. To address this issue, Brand et al [74] proposed a coupled HMM to represent the dynamics of interacting targets. They demonstrate the superiority of their approach over conventional HMMs in recognizing two-handed gestures. Incorporating domain knowledge into the HMM formalism has been investigated by several researchers. Moore et al [75] used HMMs in conjunction with object detection modules to exploit the relationship between actions and objects. Hongeng and Nevatia [76] incorporate *a priori* beliefs of state-duration into the HMM framework and the resultant model is called Hidden semi-Markov Model (semi-HMMs). Cuntoor and Chellappa [77] have proposed a mixed-state HMM formalism to model non-stationary activities, where the state-space is augmented with a discrete label for higher-level behavior modeling.

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 [69] 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 $\mathbb{R}^k$ 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)

$$x(t) = Ax(t-1) + w(t), \quad w \sim N(0, Q) \tag{1}$$
$$y(t) = Cx(t) + v(t), \quad v \sim N(0, R) \tag{2}$$

where $x \in \mathbb{R}^d$ is the d-dimensional state vector and $y \in \mathbb{R}^n$ 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. The LDS can be interpreted as a continuous state-space generalization of HMMs with a Gaussian observation model. Several applications such as recognition of humans and actions based on gait ([78], [7], [79]), activity recognition ([80]) and dynamic texture modeling and recognition [81], [82], [83], [84] have been proposed using LDSs. First order LDSs were used by Vaswani et al [9] to model the configuration of groups of people in an airport tarmac setting by considering a collection of moving points (humans) as a deforming shape.

Advances in system identification theory for learning LDS model parameters from data [85], [86], [87], [88], [81] and distance metrics on the LDS space [89], [82], [90] have made LDSs popular for learning and recognition of high-dimensional time-series data. More recently, in-depth study of the LDS space has enabled the application of machine learning tools on that space such as dynamic boosting [91], kernel methods [92], [93] and statistical modeling [94]. Newer methods to learn the model parameters [81] have made learning much more efficient than in the case of HMMs. Like HMMs, LDSs are also based on assumptions of Markovian Dynamics and conditionally independent observations. Thus, as in the case of HMMs, the time-invariant model is not applicable to non-stationary actions.

3) Non-linear Dynamical Systems: While time-invariant HMMs and LDSs are efficient modeling and learning tools, they are restricted to linear and stationary dynamics. Consider the following activity – a person bends down to pick up an object, then he walks to a nearby table and places the object on the table and finally rests on a chair. This activity is composed of a sequence of short segments each of which can be modeled as a LDS. The entire process can be seen as switching between LDSs. The most general form of the time-varying LDS is given by equations (3) and (4)

$$x(t) = A(t)x(t-1) + w(t), \quad w \sim N(0, Q) \tag{3}$$
$$y(t) = C(t)x(t) + v(t), \quad v \sim N(0, R) \tag{4}$$

which looks similar to the LDS in equations (1) and (2), except that the model parameters A and C are allowed to vary with time. To tackle such complex dynamics, a popular approach is to model the process using Switching Linear Dynamical systems (SLDS) or Jump Linear Systems (JLS). An SLDS consists of a set of LDSs with a switching function that causes model parameters to change by switching between models. Bregler [95] presented a multi-layered approach to recognize complex movements consisting of several levels of abstraction. The lowest level is a sequence of input images. The next level consists of 'blob' hypotheses where each blob

is a region of coherent motion. At the third level, blob tracks are grouped temporally. The final level, consists of a HMM for representing the complex behavior. North et al [96] augment the continuous state vector with a discrete state component to form a 'mixed' state. The discrete component represents a mode of motion or more generally a 'switch' state. Corresponding to each switch state, a Gaussian autoregressive model is used to represent the dynamics. A maximum likelihood approach is used to learn the model parameters for each motion class. Pavlovic and Rehg [97], [98] model the non-linearity in human motion in a similar framework, where the dynamics are modeled using LDS and the switching process is modeled using a probabilistic finite state-machine. Other applications of this framework include the work of Del Vecchio et al [99], [100] who used this framework for classifying drawing tasks.

Though the SLDS framework has greater modeling and descriptive power than HMMs and LDSs, learning and inference in SLDS are much more complicated, often requiring approximate methods [101]. In practice, determining the appropriate number of switching states is challenging and often requires large amounts of training data or extensive hand tuning. Apart from maximum likelihood (ML) approaches, algebraic approaches which can simultaneously estimate the number of switching states, the switching instants and also the parameters of the model for each state have been proposed by Vidal et al [102]. However, algebraic approaches are often not robust to noise and outliers in the data.

V. Modeling and Recognizing Activities

Most activities of interest in applications such as surveillance and content-based indexing involve several actors, who interact not only with each other, but also with contextual entities. The approaches discussed so far are mostly concerned with modeling and recognizing actions of a single actor. Modeling a complex scene, the inherent structure and semantics of complex activities require higher-level representation and reasoning methods. The previously discussed approaches are not suited to deal with the complexities of spatio-temporal constraints on actors and actions, temporal relations such as sequencing and synchronization, and the presence of multiple execution threads. Thus, structural and syntactic approaches such as dynamic belief networks, grammars, petri-nets etc are well-suited to tackle these problems. Moreover, some amount of domain knowledge can be exploited to design concise and intuitive structural descriptions of activities.

A. Graphical Models

1) Belief Networks: A Bayesian network (BN) [103] is a graphical model that encodes complex conditional dependencies between a set of random variables which are encoded as local conditional probability densities (CPD). Dynamic Belief networks (DBNs) are a generalization of the simpler Bayesian networks by incorporating temporal dependencies between random variables. DBNs encode more complex conditional dependence relations among several random variables as opposed to just one hidden variable as in a traditional HMM.

Huang et al. [19] used DBNs for vision based traffic monitoring. Buxton and Gong [104] used Bayesian networks to capture the dependencies between scene layout and low-level image measurements for a traffic surveillance application. Remagnino et al [105] present an approach using DBNs for scene description at two levels of abstraction — agent level descriptions and inter-agent interactions. Modeling two-person interactions such as pointing, punching, pushing, hugging etc was proposed by Park and Aggarwal [106] in a two-stage process. First, pose estimation is done via a BN and temporal evolution of pose is modeled by a DBN. Intille and Bobick [107] use Bayesian networks for multi-agent interactions where the network structure is automatically generated from the temporal structure provided by a user. Usually the structure of the DBN is provided by a domain expert. But this is difficult in real life systems where there are a very large number of variables with complex inter-dependencies. To address this issue Gong et al [108] presented a DBN framework where the structure of the network is discovered automatically using Bayesian Information Criterion [109], [110].

DBNs have also been used to recognize actions using the contextual information of the objects involved. Moore et al [75] conduct action recognition using belief networks based on scene context derived from other objects in the scene. Gupta et. al [111] present a Bayesian network for interpretation of human-object interactions that integrates information from perceptual tasks such as human motion analysis, manipulable object detection and "object reaction" determination. Filipovyich et. al [112] proposed a probabilistic graphical model of primitive actor-object interactions that combines information about the interactions dynamics, and actor-object static appearances and spatial configurations. The model is learned without any manual input of object and contextual information.

Though DBNs are more general than HMMs by considering dependencies between several random variables, the temporal model is usually Markovian as in the case of HMMs. Thus, only sequential activities can be handled by the basic DBN model. Development of efficient algorithms for learning and inference in graphical models (c. f. [113], [114]) have made them popular tools to model structured activities. Methods to learn the topology or structure of Bayesian networks from data [115] have also been investigated in the machine learning community. However, to learn the local CPDs for large networks requires very large amounts of training data or extensive hand-tuning by experts both of which limit the applicability of DBNs in large-scale settings.

2) Petri Nets: Petri Nets were defined by Carl Adam Petri [116] as a mathematical tool for describing relations between conditions and events. Petri Nets are particularly useful to model and visualize behaviors such as sequencing, concurrency, synchronization and resource sharing. Petri-nets are bi-partite graphs consisting of two types of nodes - Places and Transitions. Places refer to the state of an entity and transitions refer to changes in the state of the entity. An activity is specified by a set of entities and how the entities interact with other. Consider an example of a car pickup activity represented by a probabilistic Petri Net as shown in figure 10.

In this figure, the places are labeled $p_1, \ldots, p_5$ and transitions $t_1, \ldots, t_6$. In this PN, p_1 and p_3 are the start nodes and p_5 is the terminal node. When a car enters the scene, a 'token' is placed in place p_1. The transition t_1 is enabled in this state, but it cannot fire until the condition associated with it is satisfied i.e., when the car stops near a parking slot. When this occurs, the token is removed from p_1 and placed in p_2. Similarly when a person enters the parking lot, a token is placed in p_3 and transition t_5 fires after the person disappears near the parked car. The token is then removed from p_3 and placed in p_4. Now with a token in each of the enabling places of transition t_6, it is ready to fire when the associated condition i.e. car leaving the parking lot is satisfied. Once the car leaves, t_6 fires and both the tokens are removed and a token placed in the final place p_5. This example illustrates sequencing, concurrency and synchronization.

Petri-nets allow multiple parallel threads of execution and have traditionally found use in modeling hybrid systems, where they are well-suited to model complex behavior such as concurrency, synchronization and resource sharing [117], [118]. Petri-Nets were used by Castel et al [119] to develop a system for high-level interpretation of image sequences. In their approach the structure of the Petri-net was specified a priori. This can be tedious for large networks representing complex activities. Ghanem et al [120] proposed a method to semi-automate this task by automatically mapping a small set of logical, spatial and temporal operators to the graph structure. Using this method, they developed an interactive tool for querying surveillance videos by mapping user queries to Petri-nets. However, these approaches were based on deterministic Petri-nets. Hence they cannot deal with uncertainty in the low-level modules as is usually the case with trackers, object detectors etc. Further, real-life human activities do not conform to hard-coded models - the models need to allow deviations from the expected sequence of steps while penalizing significant deviations. To address this issue Albanese et al [121] proposed the concept of a probabilistic Petri Net (PPN) (see figure 10). In a probabilistic PN the transitions are associated with a weight which encodes the probability with which that transition fires. By using skip transitions and penalizing them with a low probability, robustness is achieved to missing observations in the input stream. Further, the uncertainty in the identity of an object or the uncertainty in the unfolding of an activity can be efficiently incorporated into the tokens of the Petri-net.

Though Petri-Nets are an intuitive tool for expressing complex activities, they suffer from the disadvantage of having to manually describe the model structure. The problem of learning the structure from training-data has not yet been formally addressed.

3) Other Graphical Models: Other graphical models have been proposed to deal with the drawbacks in DBNs - most significantly the limitation to sequential activities. Graphical models that specifically model more complex temporal relations such as sequentiality, duration, parallelism, synchrony etc have been proposed in the DBN framework. Examples include the work of Pinhanez and Bobick [122] who use a simplified version of Allen's interval algebra to model

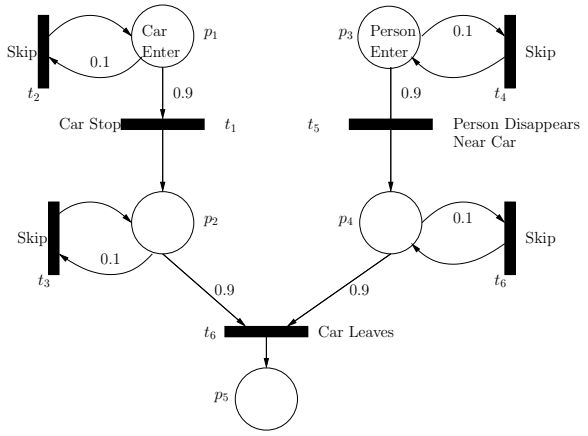

Fig. 10. A probabilistic Petri-Net representing a pickup-by-car activity. Figure taken from [121].

sophisticated temporal ordering constraints such as past, now and future. This structure is termed the PNF (past-now-future) network. Similarly, Shi et al [123], [124] have proposed using Propagation nets to represent activities using partially ordered temporal intervals. In their approach an activity is constrained by temporal and logical ordering and duration of the activity intervals. More recently, Hamid et al [125] consider a temporally extended activity as a sequence of event labels. Due to contextual and activity specific constraints the sequence labels are observed to have some inherent partial ordering. For example, in a kitchen setting the refrigerator would have to be opened before the eggs can be accessed. Using these constraints, they consider an activity model as a set of subsequences which encode the partial ordering constraints of varying lengths. These subsequences are efficiently represented using Suffix trees. The advantage of the Suffix-tree representation is that the structure of the activity can be learnt from training data using standard graph-theoretic methods.

B. Syntactic Approaches

Syntactic pattern recognition approaches based on grammars express the structure of a process using a set of production rules. To draw a parallel to grammars in language modeling, the production rules specify how complex sentences (activities) can be constructed in a grammatically sound manner from simpler words (activity primitives), and how to recognize if a given sentence (video) conforms to the rules of a given grammar (activity model). Syntactic approaches are useful when the structure of a process is difficult to learn but may be known a priori. Syntactic pattern recognition approaches were first successfully applied to still-image recognition tasks such as shape modeling [126]. Success in these domains coupled with the success of HMMs and DBNs in action-recognition tasks led to renewed interest in syntactic approaches for activity recognition.

1) Context free Grammars: One of the earliest use of grammars for visual activity recognition was proposed by Brand [127], who used a grammar to recognize hand manipulations in sequences containing disassembly tasks. They made use of simple grammars with no probabilistic modeling. Ryoo and Aggarwal [128] used the CFG formalism to model

and recognize composite human activities and multi-person interactions. They followed a hierarchical approach where the lower-levels are composed of HMMs and Bayesian Networks. The higher-level interactions are modeled by CFGs. Context-free grammar approaches present a sound theoretical basis for modeling structured processes. In syntactic approaches, one only needs to enumerate the list of primitive events that need to be detected and the set of production rules that define higher-level activities of interest. Once the rules of a grammar have been formulated, efficient algorithms to parse them exist [129], [130] which have made them popular in real-time applications.

Since deterministic grammars expect perfect accuracy in the lower-levels, they are not suited to deal with errors in low-level tasks such as tracking errors and missing observations. In complex scenarios involving several agents requiring temporal relations that are more complex than just sequencing – such as parallelism, overlap, synchrony – it is difficult to formulate the grammatical rules manually. Learning the rules of the grammar from training data is a promising alternative, but it has proved to be extremely difficult in the general case [131].

2) Stochastic Grammars: Algorithms for detection of low-level primitives are frequently probabilistic in nature. Thus, Stochastic Context-free grammars (SCFGs) which are a probabilistic extension of CFGs were found to be suitable for integration with real-life vision modules. SCFGs were used by Ivanov and Bobick [132] to model the semantics of activities whose structure was assumed to be known. They used HMMs for low-level primitive detection. The grammar production rules were augmented with probabilities and a 'skip' transition was introduced. This resulted in increased robustness to insertion errors in the input stream and also to errors in low-level modules. Moore et al [133] used SCFGs to model multi-tasked activities – activities that have several independent threads of execution with intermittent dependent interactions with each other as demonstrated in a Blackjack game with several participants. Ogale et al automatically [134] learn a Probabilistic Context-free grammar (PCFG) for human actions from sequences of human silhouettes.

In many cases, it is desirable to associate additional attributes or features to the primitive events. For example, the exact location in which the primitive event occurs may be significant for describing an event, but this may not be effectively encoded in the (finite) primitive event set. Attributes are also useful where the number of primitive events is unbounded such as in an event involving arbitrary number of objects each having distinct primitive events associated with it. Thus, attribute grammars achieve greater expressive power than traditional grammars. Probabilistic attribute grammars have been used by Joo and Chellappa [135] for multi-agent activities in surveillance settings. In the example shown in figure 11, one can see the production rules and the primitive events such as 'appear', 'disappear', 'moveclose', 'moveaway' etc in the description of the activity. The primitive events are further associated with attributes such as location (loc) where the appearance and disappearance events occur, classification (class) into a set of objects, identity (idr) of the entity involved etc.

While stochastic grammars are more robust than CFGs to errors and missed detections in the input stream, they share many of the temporal relation modeling limitations of CFGs as discussed above.

C. Knowledge and Logic-based Approaches

Logic and knowledge based approaches express activities in terms of primitives and constraints on them. These methods can express more complex constraints than grammar based approaches. While grammars can be efficiently parsed due to their syntactic structure, logical rules can lead to a computational overhead due to constraint satisfaction checks. But logical rules are often far more intuitive and human-interpretable than grammatical rules.

1) Logic Based Approaches: Logic-based methods rely on formal logical rules to describe common-sense domain knowledge to describe activities. Logical rules are useful to express domain knowledge as input by a user or to present the results of high-level reasoning in an intuitive and human-readable format. Declarative models [136] describe all expected activities in terms of scene structure, events etc. The model for an activity consists of the interactions between the objects of the scene. Medioni et al. [137] propose a hierarchical representation to recognize a series of actions performed by a single agent. Symbolic descriptors of actions are extracted from low-level features through several mid-level layers. Next, a rule based method is used to approximate the probability of occurrence of a specific activity by matching the properties of the agent with the expected distributions (represented by a mean and a variance) for a particular action. In a later work, Hongeng et al [138] extended this representation by considering an activity to be composed of several action threads. Each action thread is modeled as a stochastic finite-state automaton. Constraints between the various threads are propagated in a temporal logic network. Shet et al [139] propose a system that relies on logic programming to represent and recognize high-level activities. Low level modules are used to detect primitive events. The high level reasoning engine is based on Prolog and recognizes activities which are represented by logical rules between primitives. These approaches do not explicitly address the problem of uncertainty in the observation input stream. To address this issue, a combination of logical and probabilistic models was presented in [140] where each logical rule is represented as first-order logic formula. Each rule is further provided with a weight, where the weight indicates a belief in the accuracy of the rule. Inference is performed using a Markov-logic network.

While logic based methods are a natural way of incorporating domain knowledge, they often involve expensive constraint satisfaction checks. Further, it is not clear how much domain knowledge should be incorporated in a given setting - incorporating more and more knowledge can make the model rigid and non-generalizable to other settings. The logic rules require extensive enumeration by a domain expert for every deployment individually.

2) Ontologies: In most practical deployments that use any of the afore-mentioned approaches, symbolic activity definitions are constructed in an empirical manner, for example the

$S \rightarrow BOARDING_N$
$BOARDING \rightarrow appear_0 CHECK_1 disappear_1$
$(isPerson(appear.class) \wedge isInside(appear.loc, Gate) \wedge isInside(disappear.loc, Plane))$
$CHECK \rightarrow moveclose_0 CHECK_1$
$CHECK \rightarrow moveaway_0 CHECK_1$
$CHECK \rightarrow moveclose_0 moveaway_1 CHECK_1$
$(isPerson(moveclose.class) \wedge moveclose.idr = moveaway.idr)$

Fig. 11. Example of an attribute grammar for a passenger boarding an airplane taken from [135].

```
PROCESS(cruise-parking-lot(vehicle v, parking-lot lot),
Sequence(enter(v, lot),
   set-to-zero(i),
   Repeat-Until(
      AND(move-in-circuit(v), inside(v, lot), increment(i)),
      equal(i, n)),
   exit(v, lot)))
```

Fig. 12. Ontology for Car Cruising in Parking Lot activity. Example taken from [143].

rules of a grammar or a set of logical rules are specified manually. Though empirical constructs are fast to design and even work very well in most cases, they are limited in their utility to specific deployments for which they have been designed. Hence, there is a need for a centralized representation of activity definitions or ontologies for activities which are independent of algorithmic choices. Ontologies standardize activity definitions, allow for easy portability to specific deployments, enable interoperability of different systems and allow easy replication and comparison of system performance. Several researchers have proposed ontologies for specific domains of visual surveillance. For example, Chen et al. [141] proposed an ontology for analyzing social interaction in nursing homes, Hakeem et al for classification of meeting videos [142] and Georis et al [8] for activities in a bank monitoring setting. To consolidate these efforts and to build a common knowledge-base of domain ontologies, the Video Event Challenge Workshop was held in 2003. As a result of this workshop, ontologies have been defined for six domains of video surveillance [143] - 1) Perimeter and Internal Security, 2) Railroad Crossing Surveillance, 3) Visual Bank Monitoring, 4) Visual Metro Monitoring, 5) Store Security, 6) Airport-Tarmac Security. An example from the ontology output is shown in figure 12 which describes car cruising activity. This ontology keeps track of the number of times the car moves around in a circuit inside the parking lot without stopping. When this exceeds a set threshold, a cruising activity is detected. The workshop also led to the development of two formal languages - The Video Event Representation Language (VERL) [144], [145] which provides an ontological representation of complex events in terms of simpler sub-events and the Video Event Markup Language (VEML) which is used to annotate VERL events in videos.

Though ontologies provide concise high-level definitions of activities, they do not necessarily suggest the right 'hardware' to 'parse' the ontologies for recognition tasks.

VI. Directions for Future Work and Conclusion

A lot of enthusiasm has been generated in the vision community by recent advances in machine recognition of activities. However, several important issues remain to be addressed. In this section, we briefly discuss some of these issues.

A. Real-World Conditions

Most action and activity recognition systems are currently designed and tested on video sequences acquired in constrained conditions. Factors that can severely limit the applicability in real world conditions include noise, occlusions, shadows etc. Errors in feature extraction can easily propagate to higher-levels. For real-world deployment, action recognition systems need to be tested against such real-world conditions. Methods that are robust to these factors also need to be investigated. Many practically deployed systems do not record videos at high spatio-temporal resolution partly due to the difficulty in storing the large data that is produced. Hence, dealing with low-resolution video is an important issue. In the approaches discussed so far, it is assumed that reliable features can be extracted in a given setting such as optical-flow or background subtracted blobs. In analyzing actions in far-field settings this assumption does not usually hold. While researchers have addressed these issues in specific settings (c. f. [33], [146]), a systematic and general approach is still lacking. Hence, more research needs to be done to address these practical issues.

B. Invariances in Human Action Analysis

One of the most significant challenges in action recognition is to find methods that can explain and be robust to the wide variability in features that are observed within the same action class. Sheikh et. al. [147] have identified three important sources that give rise to variability in observed features. They are

1) Viewpoint
2) Execution Rate
3) Anthropometry

Any real-world action recognition system needs to be invariant to these factors. In this section, we will review some efforts in this direction that have been pursued in the research community.

1) View-Invariance: A fundamental problem in video-based recognition of activities is achieving view-invariant representations of actions. While it may be easy to build statistical models of simple actions based on the representations discussed so far from a single view, it is extremely challenging to generalize them to other views even for very simple action classes. This is due to the wide variations in motion based features induced by camera perspective effects and occlusions. One way to deal with the problem is to store templates from several canonical views as done in [35] and interpolate across the stored views as proposed by [148]. This approach however is not scalable since one does not know how many views to consider as canonical. Another approach is to assume that point correspondences across views are available as in [37] and compute a transformation that maps a stored model to an example from an arbitrary view. Similarly, [32] present an approach to recognize cyclic motion that is affine-invariant by assuming that feature correspondence between successive time-instants is known. It was shown by Rao and Shah [149], [150] that extrema in space-time curvature of trajectories is preserved across views which was exploited to perform view-invariant action recognition. Another example is the work of Parameswaran et al [151], [152] who define a view-invariant representation of actions based on the theory of 2D and 3D invariants. In their approach, they consider an action to be a sequence of *poses*. They assume that there exists at least one *key-pose* in the sequence in which 5 points are aligned on a plane in the 3-D world coordinates. Using this assumption, they derive a set of view-invariant descriptors. More recently, the notion of motion-history [34], [35] was extended to 3-D by Weinland et al [153] where the authors combine views from multiple cameras to arrive at a 3-D binary occupancy volume. Motion history is computed over these 3-D volumes and view-invariant features are extracted by computing circular FFT of the volume. All these approaches are strongly tied to the specific choice of feature. There is no general approach of achieving view-invariance that can be extended to several features, thus making it an open research issue.

2) Execution Rate Invariance: The second major source of observed variability in features arises from the differences in execution rates while performing the same action. Variations in execution style exist both in inter-person and intra-person settings. State-space approaches are robust to minor changes in execution rates, but are not truly rate-invariant since they do not explicitly model transformations of the temporal axis (c. f. [154], [155]). Mathematically, the variation in execution rate is modeled as a warping function of the temporal scale. The simplest case of linear time-warps can be usually dealt with fairly easily (c. f. [35], [156]). To model highly non-linear warping functions, the most common method is Dynamic Time Warping (DTW) of the feature sequence such as in [157], [148], [158], [159]. Recently, Veeraraghavan et al [160] proposed using dynamic time-warping with constraints to account for the fact that the space of all time-warp functions does not produce physically meaningful actions, hence the best time-warp has to be searched within some constraints. DTW is a promising method as it is independent of the choice of feature. The only requirement is that a distance metric be defined on the feature-space. DTW requires accurate temporal alignment of test and gallery sequences i.e. the start and end time instants have to be aligned between the test sequence and the gallery sequence. This is a strong requirement as alignment may not be accurate in a practical setting. Further, the distance computations involved are significant and can be prohibitive for long sequences involving many templates. Thus, more efficient methods are required to achieve real-time performance.

3) Anthropometric Invariance: Anthropometric variations such as those induced by the size, shape, gender etc. of humans is another important class of variabilities that requires careful attention. Unlike viewpoint and execution-rate variabilities which have received significant attention, a systematic study of anthropometric variations has been receiving interest only in recent years. Ad hoc methods which normalize the extracted features to compensate for changes in size, scale etc. are usually employed when no further information is available. Drawing on studies on human anthropometry Gritai et al [161] suggested that the anthropometric transformation between two different individuals can be modeled as a projective transformation of the image co-ordinates of body joints. Based on this, they define a similarity metric between actions by using epipolar geometry to provide constraints on actions performed by different individuals. Further research is needed in order to understand the effects of anthropometric variations and building algorithms to achieve invariance to this factor.

C. Evaluation of Complex Systems

Establishing standardized test-beds is a fundamental requirement to compare algorithms and assess progress. It is encouraging to see that several datasets have been made available by research groups and new research is expected to report results on these datasets. Examples include the UCF activity dataset [149], TSA airport tarmac dataset [9], Free Viewpoint INRIA dataset [153] and the KTH actions dataset [53]. However, most of these datasets consist of simple actions such as opening a closet door, lifting an object etc. Very few common datasets exist for evaluating higher-level complex activities and reasoning algorithms. Complex activity recognition systems consist of a slew of lower-level detection and tracking modules. Hence, a straightforward comparison of systems is not easy. One reasonable approach to evaluate complex systems is to create ground truth corresponding to outputs from a predefined set of low-level modules. Evaluation would then focus solely on the high-level reasoning engines. While this is one criteria of evaluation, the other criteria is the ability to deal with errors in low-level modules. Participation from the research community is required to address this important issue.

D. Integration with other Modalities

A vision-based system to recognize human activities can be seen as a crucial stepping stone toward the larger goal of designing machine intelligence systems. To draw a parallel with natural intelligence, humans rely on several modalities including the five classical senses – vision, audition, tactition,

olfaction, gustation – and other senses such as thermoception (temperature) and Equilibrioception (balance and acceleration) for everyday tasks. It has also been realized that alternate modalities can improve the performance of vision-based systems e.g. inertial sensors in structure-from-motion (SfM), joint audio-video based tracking [162] etc. Thus, for the longer-term pursuit to create machine intelligence, or for the shorter-term pursuit of increasing the robustness of action/activity detection modules, integration with other modalities such as audio, temperature, motion and inertial sensors needs to be investigated in a more systematic manner.

E. Intention Reasoning

Most of the approaches for recognizing and detecting action and activities are based on the premise that the action/activity has already occurred. Reasoning about the intentions of human and inferring what is going to happen presents a significant intellectual challenge. Security applications are among the first that stand to benefit from such a system, where detection of threat is of utmost importance.

Providing a machine the ability to see and understand as humans do has long fascinated scientists, engineers and even the common man. Synergistic research efforts in various scientific disciplines – Computer Vision, AI, Neuroscience, Linguistics etc – have brought us closer to this goal than at any other point in history. However, several more technical and intellectual challenges need to be tackled before we get there. The advances made so far need to consolidated, in terms of their robustness to real-world conditions and real-time performance. This would then provide a firmer ground for further research.

REFERENCES

[1] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception and Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[2] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, 1999.

[3] C. Cedras and M. Shah, "Motion-based recognition: A survey," *Image and Vision Computing*, vol. 13, no. 2, pp. 129–155, 1995.

[4] D. M. Gavrila, "The visual analysis of human movement: a survey," *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[5] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[6] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4, 2006.

[7] A. Veeraraghavan, A. Roy-Chowdhury, and R. Chellappa, "Matching shape sequences in video with an application to human movement analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1896–1909, 2005.

[8] B. Georis, M. Maziere, F. Bremond, and M. Thonnat, "A Video Interpretation Platform Applied to Bank Agency Monitoring," *2nd Workshop on Intelligent Distributed Surveillance Systems (IDSS)*, 2004.

[9] N. Vaswani, A. K. Roy-Chowdhury, and R. Chellappa, ""Shape Activity": a continuous-state HMM for moving/deforming shapes with application to abnormal activity detection," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1603–1616, 2005.

[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.

[11] 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.

[12] S. F. Chang, "The holy grail of content-based media analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 6–10, 2002.

[13] 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.

[14] 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.

[15] 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.

[16] A. Pentland, "Smart rooms, smart clothes," *International Conference on Pattern Recognition*, vol. 2, pp. 949–953, 1998.

[17] 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.

[18] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Computing Surveys*, vol. 27, no. 3, pp. 433–466, 1995.

[19] T. Huang, D. Koller, J. Malik, G. H. Ogasawara, B. Rao, S. J. Russell, and J. Weber, "Automatic symbolic traffic scene analysis using belief networks," *National Conference on Artificial Intelligence*, pp. 966–972, 1994.

[20] A. K. Roy-Chowdhury and R. Chellappa, "A factorization approach to activity recognition," *CVPR Workshop on Event Mining*, 2003.

[21] A. M. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric model for background subtraction," *Proceedings of IEEE European Conference on Computer Vision*, pp. 751–767, 2000.

[22] M.-K. Hu, "Visual pattern recognition by moment invariants," *IEEE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[23] H. Freeman, "On the encoding of arbitrary geometric configurations," *IRE Transactions on Electronic Computers*, vol. 10, no. 2, pp. 260–268, 1961.

[24] D. G. Kendall, "Shape manifolds, procrustean metrics and complex projective spaces," *Bulletin of London Mathematical society*, vol. 16, pp. 81–121, 1984.

[25] H. Blum and R. N. Nagel, "Shape description using weighted symmetric axis features," *Pattern Recognition*, vol. 10, no. 3, pp. 167–180, 1978.

[26] A. Bissacco, P. Saisan, and S. Soatto, "Gait recognition using dynamic affine invariants," *International Symposium on Mathematical Theory of Networks and Systems*, 2004.

[27] I. Laptev and T. Lindeberg, "Space-time interest points," *Proceedings of IEEE International Conference on Computer Vision*, 2003.

[28] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[29] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005.

[30] R. Polana and R. Nelson, "Low level recognition of human motion (or how to get your man without finding his body parts)," *Proceedings of the IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pp. 77–82, 1994.

[31] 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.

[32] S. M. Seitz and C. R. Dyer, "View-invariant analysis of cyclic motion," *International Journal of Computer Vision*, vol. 25, no. 3, pp. 231–251, 1997.

[33] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 781–796, 2000.

[34] J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 928–934, 1997.

[35] 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.

[36] 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.

[37] 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.

[38] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 984–989, 2005.

[39] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *Proceedings of IEEE International Conference on Computer Vision*, pp. 1395–1402, 2005.

[40] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[41] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[42] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in Neural Information Processing Systems*, pp. 585–591, 2001.

[43] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction." *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[44] 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.

[45] A. M. Elgammal and C. S. Lee, "Inferring 3D body pose from silhouettes using activity manifold learning," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 681–688, 2004.

[46] R. Pless, "Image spaces and video trajectories: Using isomap to explore video sequences," *Proceedings of IEEE International Conference on Computer Vision*, pp. 1433–1440, 2003.

[47] 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.

[48] 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.

[49] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," *Proceedings of IEEE International Conference on Computer Vision*, 2007.

[50] 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.

[51] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 123–130, 2001.

[52] J. C. Niebles, H. Wang, and L. Fei Fei, "Unsupervised learning of human action categories using spatial-temporal words," *British Machine Vision Conference*, 2006.

[53] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," *International Conference on Pattern Recognition*, pp. 32–36, 2004.

[54] S. Savarese, A. Del Pozo, J. C. Niebles, and L. Fei-Fei, "Spatial-temporal correlations for unsupervised action classification," *IEEE Workshop on Motion and Video Computing*, 2008.

[55] S. Nowozin, G. Bakir, and K. Tsuda, "Discriminative subsequence mining for action classification," *Proceedings of IEEE International Conference on Computer Vision*, pp. 1–8, 2007.

[56] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 17–31, 2007.

[57] S. F. Wong, T. K. Kim, and R. Cipolla, "Learning motion categories using both semantic and structural information," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–6, 2007.

[58] Y. Song, L. Goncalves, and P. Perona, "Unsupervised learning of human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 814–827, 2003.

[59] J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[60] E. Shechtman and M. Irani, "Space-time behavior based correlation," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 405–412, 2005.

[61] P. A. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[62] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," *Proceedings of IEEE International Conference on Computer Vision*, pp. 166–173, 2005.

[63] ——, "Spatio-temporal shape and flow correlation for action recognition," *Visual Surveillance Workshop*, 2007.

[64] M. A. O. Vasilescu, "Human motion signatures: Analysis, synthesis, recognition," *International Conference on Pattern Recognition*, pp. 456–460, 2002.

[65] T. K. Kim, S. F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[66] L. Wolf, H. Jhuang, and T. Hazan, "Modeling appearances with low-rank svm," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[67] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[68] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 379–385, 1992.

[69] J. Schlenzig, E. Hunter, and R. Jain, "Recursive identification of gesture inputs using hidden markov models," *Proceedings of the Second IEEE Workshop on Applications of Computer Vision*, pp. 187–194, 1994.

[70] A. D. Wilson and A. F. Bobick, "Learning visual behavior for gesture analysis," *Proceedings of the International Symposium on Computer Vision*, pp. 229–234, 1995.

[71] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer based video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371–1375, 1998.

[72] A. Kale, A. Sundaresan, A. N. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger, and R. Chellappa, "Identification of humans using gait," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1163–1173, 2004.

[73] Z. Liu and S. Sarkar, "Improved gait recognition by gait dynamics normalization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 863–876, 2006.

[74] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 994–999, 1997.

[75] D. J. Moore, I. A. Essa, and M. H. Hayes, "Exploiting human actions and object context for recognition tasks," *Proceedings of IEEE International Conference on Computer Vision*, pp. 80–86, 1999.

[76] S. Hongeng and R. Nevatia, "Large-scale event detection using semi-hidden markov models," *Proceedings of IEEE International Conference on Computer Vision*, pp. 1455–1462, 2003.

[77] N. P. Cuntoor and R. Chellappa, "Mixed-state models for nonstationary multiobject activities," *EURASIP Journal of Applied Signal Processing*, vol. 2007, no. 1, pp. 106–119, 2007.

[78] A. Bissacco, A. Chiuso, Y. Ma, and S. Soatto, "Recognition of human gaits," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 52–57, 2001.

[79] M. C. Mazzaro, M. Sznaier, and O. Camps, "A model (in)validation approach to gait classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1820–1825, 2005.

[80] N. P. Cuntoor and R. Chellappa, "Epitomic representation of human activities," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[81] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *International Journal of Computer Vision*, vol. 51, no. 2, pp. 91–109, 2003.

[82] P. Saisan, G. Doretto, Y. N. Wu, and S. Soatto, "Dynamic texture recognition," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 58–63, 2001.

[83] A. B. Chan and N. Vasconcelos, "Classifying video with kernel dynamic textures," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[84] ——, "Mixtures of dynamic textures," *Proceedings of IEEE International Conference on Computer Vision*, pp. 641–647, 2005.

[85] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the em algorithm," *Journal of Time Series Analysis*, vol. 3, no. 4, pp. 253–264, 1982.

[86] P. V. Overschee and B. D. Moor, "Subspace algorithms for the stochastic identification problem," *Automatica*, vol. 29, no. 3, pp. 649–660, 1993.

[87] Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," Department of Computer Science, University of Toronto, Technical Report, Tech. Rep. CRG-TR-96-2, 1996.

[88] L. Ljung, Ed., *System identification (2nd ed.): theory for the user*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1999.

[89] K. D. Cock and B. D. Moor, "Subspace angles between arma models," *Systems and Control Letters*, vol. 46, pp. 265–270, 2002.

[90] R. J. Martin, "A metric for arma processes," *IEEE Transactions on Signal Processing*, vol. 48, no. 4, pp. 1164–1170, 2000.

[91] R. Vidal and P. Favaro, "Dynamicboost: Boosting time series generated by dynamical systems," *Proceedings of IEEE International Conference on Computer Vision*, 2007.

[92] S. V. N. Vishwanathan, A. J. Smola, and R. Vidal, "Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes," *International Journal of Computer Vision*, vol. 73, no. 1, pp. 95–119, 2007.

[93] A. Bissacco and S. Soatto, "On the blind classification of time series," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[94] P. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical analysis on stiefel and grassmann manifolds with applications in computer vision," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[95] C. Bregler, "Learning and recognizing human dynamics in video sequences," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, p. 568, 1997.

[96] B. North, A. Blake, M. Isard, and J. Rittscher, "Learning and classification of complex dynamics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 9, pp. 1016–1034, 2000.

[97] V. Pavlovic, J. M. Rehg, and J. MacCormick, "Learning switching linear models of human motion," *Advances in Neural Information Processing Systems*, pp. 981–987, 2000.

[98] V. Pavlovic and J. M. Rehg, "Impact of dynamic model learning on classification of human motion," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1788–1795, 2000.

[99] D. Del Vecchio, R. M. Murray, and P. Perona, "Primitives for human motion: A dynamical approach," *Proceedings of IFAC World Congress*, 2002.

[100] ——, "Decomposition of human motion into dynamics based primitives with application to drawing tasks," *Automatica*, vol. 39, pp. 2085–2098, 2003.

[101] S. M. Oh, J. M. Rehg, T. R. Balch, and F. Dellaert, "Data-driven mcmc for learning and inference in switching linear dynamic systems," *National Conference on Artificial Intelligence*, pp. 944–949, 2005.

[102] R. Vidal, A. Chiuso, and S. Soatto, "Observability and identifiability of jump linear systems," *Proceedings of IEEE Conference on Decision and Control*, pp. 3614–3619, 2002.

[103] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[104] H. Buxton and S. Gong, "Visual surveillance in a dynamic and uncertain world," *Artificial Intelligence*, vol. 78, no. 1-2, pp. 431–459, 1995.

[105] P. Remagnino, T. Tan, and K. Baker, "Agent orientated annotation in model based visual surveillance," *Proceedings of IEEE International Conference on Computer Vision*, pp. 857–862, 1998.

[106] S. Park and J. K. Aggarwal, "Recognition of two-person interactions using a hierarchical bayesian network," *ACM Journal of Multimedia Systems, Special Issue on Video Surveillance*, vol. 10, no. 2, pp. 164–179, 2004.

[107] S. S. Intille and A. F. Bobick, "A framework for recognizing multi-agent action from visual evidence," *National Conference on Artificial Intelligence*, pp. 518–525, 1999.

[108] S. Gong and T. Xiang, "Recognition of group activities using dynamic probabilistic networks," *Proceedings of IEEE International Conference on Computer Vision*, pp. 742–749, 2003.

[109] R. L. Kashyap, "Bayesian comparison of different classes of dynamic models using empirical data," *IEEE Transactions on Automatic Control*, vol. 22, no. 5, pp. 715–727, 1977.

[110] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[111] A. Gupta and L. S. Davis, "Objects in action: An approach for combining action understanding and object perception," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[112] R. Filipovych and E. Ribeiro, "Recognizing primitive interactions by exploring actor-object states," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[113] M. I. Jordan, *Learning in Graphical Models*. The MIT Press, 1998.

[114] F. R. Kschischang, B. J. Frey, and H. A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[115] N. Friedman and D. Koller, "Being Bayesian about Bayesian network structure: A Bayesian approach to structure discovery in Bayesian networks." *Machine Learning*, vol. 50, no. 1–2, pp. 95–125, 2003.

[116] C. A. Petri, "Communication with automata," *DTIC Research Report AD0630125*, 1966.

[117] R. David and H. Alla, "Petri nets for Modeling of Dynamic Systems A Survey," *Automatica*, vol. 30, no. 2, pp. 175–202, 1994.

[118] T. Murata, "Petri nets: Properties, Analysis and Applications," *Proceedings of the IEEE*, vol. 77, no. 4, pp. 541–580, 1989.

[119] C. Castel, L. Chaudron, and C. Tessier, "What is going on? A High-Level Interpretation of a Sequence of Images," *ECCV Workshop on Conceptual Descriptions from Images*, 1996.

[120] N. Ghanem, D. DeMenthon, D. Doermann, and L. Davis, "Representation and Recognition of Events in Surveillance Video Using Petri Nets," *Second IEEE Workshop on Event Mining 2004, CVPR2004*, 2004.

[121] M. Albanese, R. Chellappa, V. Moscato, A. Picariello, V. S. Subrahmanian, P. Turaga, and O. Udrea, "A constrained probabilistic petri net framework for human activity detection in video," *Accepted for publication in IEEE Transactions on Multimedia*.

[122] C. S. Pinhanez and A. F. Bobick, "Human action detection using pnf propagation of temporal constraints," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, p. 898, 1998.

[123] Y. Shi, Y. Huang, D. Minen, A. Bobick, and I. Essa, "Propagation networks for recognizing partially ordered sequential action," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 862–869, 2004.

[124] Y. Shi, A. F. Bobick, and I. A. Essa, "Learning temporal sequence model from partially labeled data," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1631–1638, 2006.

[125] R. Hamid, A. Maddi, A. Bobick, and I. Essa, "Structure from statistics - unsupervised activity analysis using suffix trees," *Proceedings of IEEE International Conference on Computer Vision*, 2007.

[126] K. S. Fu, *Syntactic Pattern Recognition and Applications*. Prentice-Hall Inc., 1982.

[127] M. Brand, "Understanding manipulation in video," *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, p. 94, 1996.

[128] M. S. Ryoo and J. K. Aggarwal, "Recognition of composite human activities through context-free grammar based representation," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1709–1718, 2006.

[129] J. Earley, "An efficient context-free parsing algorithm," *Communications of the ACM*, vol. 13, no. 2, pp. 94–102, 1970.

[130] A. V. Aho and J. D. Ullman, *The Theory of Parsing, Translation, and Compiling, Volume 1: Parsing*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

[131] C. D. L. Higuera, "Current trends in grammatical inference," *Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition*, pp. 28–31, 2000.

[132] Y. A. Ivanov and A. F. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 852–872, 2000.

[133] D. Moore and I. Essa, "Recognizing multitasked activities from video using stochastic context-free grammar," *Eighteenth national conference on Artificial intelligence*, pp. 770–776, 2002.

[134] A. S. Ogale, A. Karapurkar, and Y. Aloimonos, "View-invariant modeling and recognition of human actions using grammars," *ECCV Workshop on Dynamical Vision*, pp. 115–126, 2006.

[135] S. W. Joo and R. Chellappa, "Recognition of multi-object events using attribute grammars," *International Conference on Image Processing*, pp. 2897–2900, 2006.

[136] N. Rota and M. Thonnat, "Activity recognition from video sequences using declarative models," *Proceedings of the 14th European Conference on Artificial Intelligence*, pp. 673–680, 2000.

[137] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 873–889, 2001.

[138] S. Hongeng, R. Nevatia, and F. Bremond, "Video-based event recognition: activity representation and probabilistic recognition methods," *Computer Vision and Image Understanding*, vol. 96, no. 2, pp. 129–162, 2004.

[139] V. D. Shet, D. Harwood, and L. S. Davis, "Vidmap: video monitoring of activity with prolog," *Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 224–229, 2005.
[140] S. Tran and L. S. Davis, "Visual event modeling and recognition using markov logic networks," *Proceedings of IEEE European Conference on Computer Vision*, 2008.
[141] D. Chen, J. Yang, and H. D. Wactlar, "Towards Automatic Analysis of Social Interaction Patterns in a Nursing Home Environment from Video," *MIR 04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pp. 283–290, 2004.
[142] A. Hakeem and M. Shah, "Ontology and Taxonomy Collaborated Framework for Meeting Classification." *International Conference on Pattern Recognition*, pp. 219–222, 2004.
[143] Event Ontology Workshop. http://www.ai.sri.com/∼burns/EventOntology.
[144] J. Hobbs, R. Nevatia, and B. Bolles, "An Ontology for Video Event Representation," *IEEE Workshop on Event Detection and Recognition*, 2004.
[145] A. R. J. Francois, R. Nevatia, J. Hobbs, and R. C. Bolles, "Verl: An ontology framework for representing and annotating video events," *IEEE Transactions on MultiMedia*, vol. 12, no. 4, pp. 76–86, 2005.
[146] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," *Proceedings of IEEE International Conference on Computer Vision*, pp. 726–733, 2003.
[147] Y. Sheikh, M. Sheikh, and M. Shah, "Exploring the space of a human action," *Proceedings of IEEE International Conference on Computer Vision*, pp. 144–149, 2005.
[148] T. J. Darrell, I. A. Essa, and A. P. Pentland, "Task-specific gesture analysis in real-time using interpolated views," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 12, pp. 1236–1242, 1996.
[149] C. Rao and M. Shah, "View-invariance in action recognition," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 316–322, 2001.
[150] C. Rao, A. Yilmaz, and M. Shah, "View-invariant representation and recognition of actions," *International Journal of Computer Vision*, vol. 50, no. 2, pp. 203–226, 2002.
[151] V. Parameswaran and R. Chellappa, "View invariants for human action recognition," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 613–619, 2003.
[152] ——, "View invariance for human action recognition," *International Journal of Computer Vision*, vol. 66, no. 1, 2006.
[153] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 249–257, 2006.
[154] A. F. Bobick and A. D. Wilson, "A state-based technique for the summarization and recognition of gesture," *Proceedings of IEEE International Conference on Computer Vision*, pp. 382–388, 1995.
[155] J. Hoey and J. J. Little, "Representation and recognition of complex human motion," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1752–1759, 2000.
[156] P. K. Turaga, A. Veeraraghavan, and R. Chellappa, "From videos to verbs: Mining videos for activities using a cascade of dynamical systems," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
[157] K. Takahashi, S. Seki, E. Kojima, and R. Oka, "Recognition of dexterous manipulations from time-varying images," *Proceedings IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pp. 23–28, 1994.
[158] M. A. Giese and T. Poggio, "Morphable models for the analysis and synthesis of complex motion patterns," *International Journal of Computer Vision*, vol. 38, no. 1, pp. 59–73, 2000.
[159] C. Rao, M. Shah, and T. Syeda-Mahmood, "Invariance in motion analysis of videos," *Proceedings of the eleventh ACM International Conference on Multimedia*, pp. 518–527, 2003.
[160] A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury, "The function space of an activity," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 959–968, 2006.
[161] A. Gritai, Y. Sheikh, and M. Shah, "On the use of anthropometry in the invariant analysis of human actions," *International Conference on Pattern Recognition*, pp. 923–926, 2004.
[162] V. Cevher, A. Sankaranarayanan, J. H. McClellan, and R. Chellappa, "Target tracking using a joint acoustic video system," *IEEE Transactions on Multimedia*, vol. 9, no. 4, pp. 715–727, 2007.