VRIJE UNIVERSITEIT BRUSSEL
FACULTY OF ENGINEERING
Department of Electronics and Informatics (ETRO)
Image Processing and Machine Vision Group (IRIS)

Visual Attention Framework: Application to Event Analysis

Thesis submitted in fulfilment of the requirements for the award of the degree of
Doctor in de ingenieurswetenschappen (Doctor in Engineering)
by

Geerinck Thomas

Examining committee
Prof. Hichem Sahli, Vrije Universiteit Brussel, promoter
Dr. Valentin Enescu, Vrije Universiteit Brussel, co-promoter
Prof. Ann Nowé, Vrije Universiteit Brussel
Prof. Rik Pintelon, Vrije Universiteit Brussel
Prof. Werner Verhelst, Vrije Universiteit Brussel
Prof. Eric Soetens, Vrije Universiteit Brussel
Prof. Jan Cornelis, Vrije Universiteit Brussel
Prof. Robert B. Fisher, University of Edinburgh
Dr. Lucas Paletta, Institute of Digital Image Processing

Address
Pleinlaan 2, B-1050 Brussels, Belgium
Tel: +32-2-6291300
Fax: +32-2-6292883
Email: tgeerinc@etro.vub.ac.be

Brussels, May 2009

Contents

List of Figures

List of Tables

Abstract

One of the monolithic goals of computer vision is to automatically interpret general digital images or videos of arbitrary scenes. However, the amount of visual information available to a vision system is enormous and in general it is computationally impossible to process all of this information bottom-up. To ensure that the process has tractable computational properties, visual attention plays a crucial role in terms of selection of visual information, allowing monitoring objects or regions of visual space and select information from them for report, recognition, etc. This dissertation discusses one small but critical slice of a cognitive computer vision system, that of visual attention. In contrast to the attention mechanisms used in most previous machine vision systems, which drive attention based on the spatial location hypothesis, in this work we propose a novel model of object-based visual attention, in which the mechanisms which direct visual attention are object-driven. Considering the temporal dynamics associated with attention-dependent motion, an attention-based visual motion framework is also proposed. Finally, since a vision system will always have a set of tasks that defines the purpose of the visual process, a top-down approach is proposed to define the competition of the visual attention occurring not only within an object but also between objects, and illustrated in the framework of a surveillance system.

Chapter 1

Semantic Video Interpretation Framework

1.1 Introduction

While event detection and localization of conspicuous visual events is indispensable, recognizing and understanding visual behavior is essential for complete dynamic scene analysis in applications such as visual surveillance and monitoring [76, 71], yet unsolved problems. One of the essential tasks for an automated vision system is to detect conspicuous, or informative activity, among other activities. By autonomous events, we imply that both the number of meaningful events and their whereabouts in the scene are automatically detected and localized in the scene, rather than manually labeled.

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 in such a way that it can automatically recognize what activity is being or was performed? Several survey papers have appeared reporting on research on machine recognition of human activities. Most notably among them are the following. Aggarwal and Cai [1] discuss three important subproblems that together form a complete action recognition system - extraction of human body structure from images, tracking across frames, and action recognition. Cedras and Shah [9] present a survey on motion-based approaches for 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 [21] presented a survey focused mainly on tracking of hands and humans via 2-D or 3-D models and a discussion of action recognition techniques. More recently, Moeslund et al. [47] presented a survey of problems and approaches in human motion capture including human model initialization, tracking, pose estimation, and activity recognition. These surveys discuss thoroughly lower level modules of detection and tracking. Recently, Turaga et al. [72] present a survey focusing exclusively on approaches for recognition of action and activities from video.

Machine vision-based activity recognition systems typically follow a hierarchical approach. At the lower levels are modules such as background-foreground segmentation, tracking and object detection. At the midlevel are primitive action recognition modules. At the high level are the reasoning engines that encode the activity semantics based on the previous levels' action primitives. It is clear that we need to define and distinguish between (primitive) action and activity.

The terms "action" and "activity" are frequently used interchangeably in the computer vision literature. The term "action" refers to simple motion patterns usually executed by a single person and typically lasting for a short duration of time. On the other hand, "activity" refers to the complex sequence of actions performed by several humans who could be interacting in a constrained manner. They are typically characterized by much longer temporal durations. In [72], these definitions conceptualize two levels of complexity to study the problem and provide a starting point to organize the numerous approaches that have been proposed.

It is clear that this is not the only possible viewpoint on human activity. Defining concepts such as "action" and "activity" is a direct result from the straightforward classification of events as being "simple" or being "complex". However, one may note that simple and complex events are not separated by a hard boundary. There is a significant gray area between these two extremes, as an event that is primitive from one perspective can be complex from another one. We can regard walking from one point to another as a simple action in a larger activity. But from another point of view, walking is a complex action consisting of repetitions of moving one leg forward and then moving the other leg forward, and the decomposition can continue even more. However, the level of decomposition is also limited by the detection algorithms, and their level of precision in generating events.

1.1.1 Event Analysis: State of the Art

In the literature, a variety of approaches have been proposed for the detection of events in video sequences. Moreover, this multitude of approaches has been categorized in several ways. In this work, we tend to follow the proposed categorization of approaches for modeling human activity in video of [72] and [29]. Most of these approaches can be arranged into two categories based on the semantic significance of their representations. This distinction is important, since it determines whether humans can exploit the representation for communication. Approaches where representations do not take into account semantic meaning do not lend themselves directly to interpretation or interface to humans.

Following the earlier proposed conceptual distinction between "action" and "activity", Figure 1.1 overviews the possible approaches.

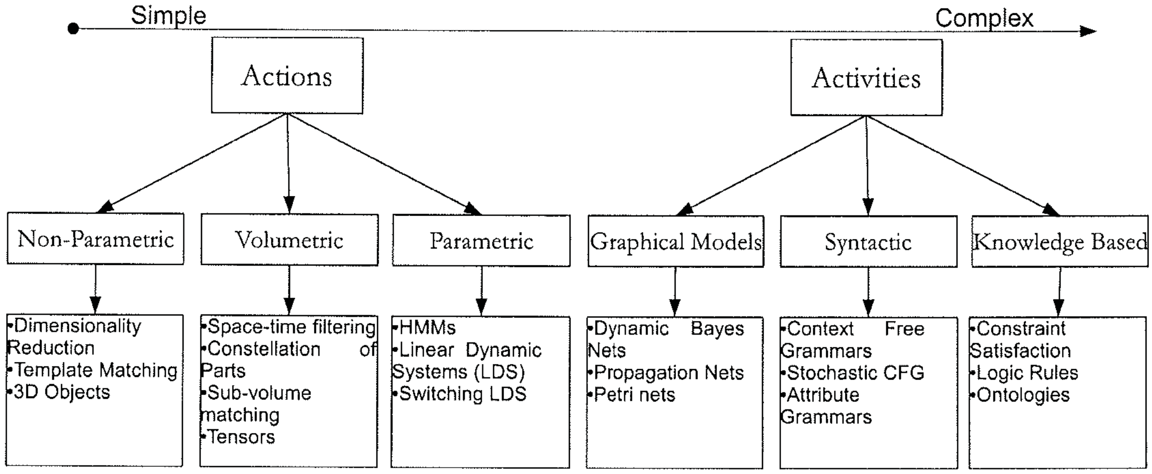

Figure 1.1: Overview of approaches for action and activity recognition. Figure taken from [72]

Approaches for modeling actions are categorized into three major classes - nonparametric, volumetric, and parametric time-series approaches.

Nonparametric approaches

Nonparametric approaches typically extract a set of features from each frame of the video. The features are then matched to a stored template. For example,

- [5, 6] propose "temporal templates" as models for actions, by aggregating subtracted motion blobs into a single static image. As such, "motion energy image" (MEI) (equal weights to all frames in sequence) and "motion history image" (MHI) (decaying weights to images in sequence with the highest weight for the most recent image) comprise together a template for a given action. From the templates, translation, rotation, and scale invariant Hu moments [34] are extracted and used for recognition. For several simple action classes, such as "sitting down", "bending", and other aerobic postures, MEI and MHI possess sufficient discriminative power. However, for complex activities, this discriminative ability is lost due to overwriting the motion history.
- [78] represents 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.

Similar to this approach, in [25] proposed a stack of blobs instead of contours which create an (x, y, t) binary space-time (ST) volume. From this ST volume 3-D shape descriptors are extracted. Because these approaches require careful segmentation of background and foreground, they are limited in applicability to fixed camera settings.

Volumetric approaches

Volumetric approaches consider video as a 3-D volume of pixel intensities and extend standard image features such as scale-space extrema, spatial filter responses etc. to the 3-D case.

- Spatiotemporal filtering approaches [14, 79] 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.

- 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. [43] proposed a spatiotemporal generalization of the well-known Harris interest point detector, widely used in object recognition. Similar approaches are [16, 51]. In most of these approaches, the detection of the parts is based on linear operations such as filtering and spatio-temporal gradients, hence the descriptors are sensitive to changes in appearance, noise, occlusions, etc.

- As opposed to part-based approaches, researchers have also investigated matching of videos by matching subvolumes between a video and a template. Inspired by the success of Haar-type features or ”box features” in object detection [75], [42] extended this framework to 3-D. Subvolume approaches are susceptible to changing backgrounds, but are more robust to noise and occlusions.

Parametric time-series approaches

Parametric time-series approaches include hidden Markov models (HMMs), linear dynamical systems (LDSs), etc. The parametric approaches are better suited for more complex actions that are temporally extended.

- 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. An excellent detailed explanation of HMMs and its associated three problems - inference, decoding, and learning - can be found in [60]. HMMs have been successfully applied for gesture recognition [77, 68], human gait [40, 44], interacting agents performing an action (coupled HMMs) [7]. 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 or when accurate start and end times for action units are unknown. However, their utility is restricted to relatively simple and stationary temporal patterns, most significantly, due to the assumption of Markovian dynamics and the time-invariant nature of the model.

- 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, described by

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

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 $d \ll n$. w and v are the process and observation noise, respectively, which are Gaussian distributed with zero-means and covariance matrices Q and R, respectively. Examples found in literature using LDSs are [45, 10, 74, 4]. Like HMMs, LDSs are also based on assumptions of Markovian dynamics and conditionally independent observations, and as such not applicable to nonstationary actions.

- While time-invariant HMMs and LDSs are efficient modeling and learning tools, they are restricted to linear and stationary dynamics. The most general form of a non-linear (time-varying) LDS is given by

  $$\begin{aligned} x(t) &= A(t)x(t-1) + w(t), \quad w \sim N(0, Q) \\ y(t) &= C(t)x(t-1) + v(t), \quad v \sim N(0, R) \end{aligned}$$

  where 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 (SLDSs) or jump linear systems (JLSs). Approaches in literature are [52, 55, 53]. 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. In practice, determining the appropriate number of switching states is challenging and often requires large amounts of training data or extensive hand tuning.

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.

Graphical Models

- Belief Networks: A Bayesian network (BN) [56] is a graphical model that encodes complex conditional dependencies between a set of random variables that are encoded as local conditional probability densities (CPD). Dynamic belief networks (DBNs) are a generalization of the simpler BNs 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. Examples in literature are [8, 54, 36]. Usually the structure of the DBN is provided by a domain expert. However, this is difficult in real-life systems where there are a very large number of variables with complex interdependencies. To address this issue, [24] presented a DBN framework where the structure of the network is discovered automatically using Bayesian information criterion [41, 64]. 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 (cf., [39]) have made them popular tools to model structured activities. Methods to learn the topology or structure of BNs from data [20] 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.

- Petri Nets: Petri nets were defined by Petri [58] 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 [15, 49]. Petri nets are applied in [23] for querying surveillance videos by mapping user queries to Petri nets. However, these approaches are based on deterministic Petri nets. In order to deal with uncertainty in low-level modules as is usually the case with trackers and object detectors, and with allowed deviations from the expected sequence steps in real-life human activity, the concept of probabilistic Petri net (PPN) is proposed in [3]. 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 been formally addressed yet.

- 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 past-now-future (PNF) network [59], propagation nets using partially ordered temporal intervals [67, 66], modeling activity as subsequences of event labels represented by Suffix-trees [30].

Syntactic Approaches

- Grammars: Grammars express the structure of a process using a set of production rules. To draw a parallel to grammars in language modeling, production rules specify how sentences (activities) can be constructed from words (activity primitives), and how to recognize if a sentence (video) conforms to the rules of a given grammar (activity model). The context-free grammar (CFG) formalism to model and recognize composite human activities and multiperson interactions is introduced in [62]. Once the rules of a CFG have been formulated, efficient algorithms to parse them exist [17, 2], which have made them popular in real-time applications. Because 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 [31].

- 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, are suitable for integration with real-life vision modules. SCFGs are used in [37, 48]. In many cases, it is desirable to associate additional attributes or features to the primitive events. Probabilistic attribute grammars, with greater expressive power than traditional grammars, have been used in [38]. While SCFGs 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.

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 [61] 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. [46] 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. [33] 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. [65] 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 [70], 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 knowledge can potentially make the model rigid and nongeneralizable to other settings. Further, the logic rules require extensive enumeration by a domain expert for every deployment.

Knowledge-Based Approaches: Ontologies

In most practical deployments that use any of the aforementioned approaches, symbolic activity definitions are constructed in an empirical manner, for example, the 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 that 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.

Ontologies are tools for structuring knowledge [11]. An ontology may be defined as the specification of a representation vocabulary for a shared domain of discourse which may include definitions of classes, relations, functions and other objects [26]. The terms of the ontology structure are called meta-concepts (e.g. event) and their instances (e.g. the ”stand-up” event) are the concepts for a particular ontology.

Definition. An object ontology is a structure

$$\mathcal{O} := (\mathcal{D}, \leq_{\mathcal{D}}, \mathcal{R}, \sigma, \leq_{\mathcal{R}})$$

consisting of: (i) Two disjoint sets D and R whose elements d and r are called respectively, intermediate level descriptors (e.g. intensity, position, etc.) and relation identifiers (e.g. relative position). To simplify the terminology, relation identifiers will often be called *relations* in the sequel. The elements of set D are often called *concept identifiers* or *concepts* in the literature. (ii) A partial order $\leq_D$ on D, called concept hierarchy or taxonomy (e.g. luminance is a subconcept of intensity). (iii) A function $\sigma : R \rightarrow D^+$ called *signature*; $\sigma(r) = (\sigma_{1,r}, \sigma_{2,r}, ...\sigma_{\sum,r})$, $\sigma_{i,r} \in D$ and $|\sigma(r)| \equiv \sum$ denotes the number of elements of D on which $\sigma(r)$ depends. (iv) A partial order $\leq_R$ on R, called relation hierarchy, where $r1 \leq_R r2$ implies $|\sigma(r1)| = |\sigma(r2)|$ and $\sigma_{i,r1} =_D \sigma_{i,r2}$ for each $1 \leq i \leq |\sigma(r1)|$.

For example, the signature of relation r relative position is by definition $\sigma(r) =$ (position, position), indicating that it relates a position to a position; $|\sigma(r)| = 2$ denotes that r involves two elements of set D. Both the intermediate-level position descriptor values and the underlying low-level descriptor values can be employed by the relative position relation.

Several researchers have proposed ontologies for specific domains of visual surveillance. For example, Chen et al. [13] proposed an ontology for analyzing social interaction in nursing homes, Hakeem et al. for classification of meeting videos [28], and Georis et al. [22] 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 [27]: 1) perimeter and internal security; 2) railroad crossing surveillance; 3) visual bank monitoring; 4) visual metro monitoring; 5) store security; and 6) airport-tarmac security. The workshop also led to the development of two formal languages-the video event representation language (VERL) [32, 19], which provides an ontological representation of complex events in terms of simpler subevents, 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.

1.1.2 Application Domains

The objective of automated video interpretation is to understand and recognize automatically behavior evolved in the observed scene. In this section we present shortly a few application areas of vision-based activity recognition systems.

Behavioral Biometrics Biometrics involve study of approaches and algorithms for uniquely recognizing humans based on physical or behavioral cues. The advantage of using behavior as cue is that subject cooperation is not necessary and it can proceed without interrupting or interfering with the subject's activity. Currently, the most promising example of behavioral biometrics is human gait [63].

Content-Based Video Analysis 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 and retrieval of consumer content such as sports videos is one of the most commercially viable applications of this technology [12].

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. As a consequence, security agencies are seeking vision-based solutions to assist the tasks of a human operator. Automatic recognition of anomalies in a camera's field of view, and querying activities of interest in a large database by learning patterns of activity from long videos are problems that attracted attention from vision researchers [73, 80, 69, 35].

Interactive Applications and Environments Understanding the interaction between computer and human remains one of the enduring challenges in designing human-computer interfaces. Effective utilization of visual cues for nonverbal communication, such as gestures, can augment this interaction. Similarly, interactive environments such as smart rooms [57], that can react on user's gestures can benefit from vision-based methods.

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 in the other hand has traditionally relied more on human animation to provide high-quality animation. However, with improvements in algorithms and hardware [18], much more realistic motion synthesis can be achieved. A related application is learning in simulated environments, including training of military soldiers or firefighters.

Bibliography

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Comput. Vis. Image Understand.*, 73(3):428–440, 1999.

[2] A. V. Aho and J. D. Ullman. *The Theory of Parsing, Translation, and Compiling.* Prentice-Hall, Englewood Cliffs, NJ, 1972.

[3] 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. *IEEE Trans. Multimedia*, to be published.

[4] A. Bissacco and S. Soatto. On the blind classification of time series. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1–7, 2007.

[5] A. F. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. *Philosoph. Trans. Roy. Soc. Lond. B*, 352:1257–1265, 1997.

[6] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(3):257–267, Mar. 2001.

[7] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 994–999, 1997.

[8] H. Buxton and S. Gong. Visual surveillance in a dynamic and uncertain world. *Artif. Intell.*, 78(1-2):431–459, 1995.

[9] C. Cedras and M. Shah. Motion-based recognition: A survey. *Image Vis. Comput.*, 13(2):129–155, 1995.

[10] A. B. Chan and N. Vasconcelos. Classifying video with kernel dynamic textures. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1–6, 2007.

[11] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins. What are ontologies and why do we need them. *IEEE Intelligent Systems*, 14:20–26, Jan. Feb. 1999.

[12] S. F. Chang. The holy grail of content-based media analysis. *IEEE Multimedia Mag.*, 9(2):6–10, Apr. 2002.

[13] D. Chen, J. Yang, and H. D. Wactlar. Towards automatic analysis of social interaction patterns in a nursing home environment from video. In *Proc. 6th ACM SIGMM Int. Workshop Multimedia Inf. Retrieval*, pages 283–290, 2004.

[14] O. Chomat and J. L. Crowley. Probabilistic recognition of activity using local appearance. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, volume 02, pages 104–109, 1999.

[15] R. David and H. Alla. Petri nets for modeling of dynamic systems: a survey. *Automatics*, 30(2):175–202, 1994.

[16] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proc. IEEE Int. Workshop Vis. Surveillance Performance Eval. Tracking Surveillance*, pages 65–72, 2005.

[17] J. Earley. An efficient context-free parsing algrithm. *Commun. ACM*, 13(2):94–102, 1970.

[18] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan. Computational studies of human motion: Part 1, tracking and motion synthesis. *Found. Trends Comput. Graphics Vis.*, 1(2-3):77–254, 2005.

[19] A. R. J. Francois, R. Nevatia, J. Hobbs, and B. R. C. Verl: An ontology framework for representing and annotating video events. *IEEE MultiMedia Mag.*, 12(4):76–86, Oct.-Dec. 2005.

[20] N. Friedman and D. Koller. Being bayesian about bayesian network structure: A bayesian approach to structure discovery in bayesian networks. *Mach. Learn.*, 50(1-2):95–125, 2003.

[21] D. M. Gavrila. The visual analysis of human movement: A survey. *Comput. Vis. Image Understand.*, 73(1):82–98, 1999.

[22] B. Georis, M. Maziere, F. Bremond, and M. Thonnat. A video interpretation platform applied to bank agency monitoring. In *Proc. 2nd Workshop Intell. Distributed Surveillance Syst.*, pages 46–50, 2004.

[23] N. Ghanem, D. DeMenthon, D. Doermann, and L. Davis. Representation and recognition of events in surveillance video using petri nets. In *Proc. 2nd IEEE Workshop Event Mining*, page 112, 2004.

[24] S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. In *Proc. IEEE Conf. Comput. Vis.*, pages 742–749, 2003.

[25] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(12):2247–2253, Dec. 2007.

[26] T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220, 1993.

[27] S. Guler, J. B. Burns, A. Hakeem, Y. Sheikh, M. Shah, M. Thonnat, F. Bremond, N. Maillot, T. V. Vu, I. Haritaoglu, R. Chellappa, U. Akdemir, and L. Davis. An ontology of video events in the physical security and surveillance domain. Online, 2003. work done as part of the ARDA video event Challenge Workshop.

[28] A. Hakeem and M. Shah. Ontology and taxonomy collaborated framework for meeting classification. In *Proc. Int. Conf. Pattern Recognit.*, pages 219–222, 2004.

[29] A. Hakeem, Y. Sheikh, and M. Shah. Case^e: A hierarchical event representation for the analysis of videos. In *Proc. of American Association of Artificial Intelligence (AAAI)*, pages 263–268, 2004.

[30] R. Hamid, A. Maddi, A. Bobick, and I. Essa. Structure from statistics - unsuservised activity analysis using suffix trees. In *Proc. IEEE Conf. Comput. Vis.*, pages 1–8, 2007.

[31] C. D. L. Higuera. Current trends in grammatical inference. In *Proc. Joint IAPR Int. Workshops Adv. Pattern Recognit.*, pages 28–31, 2000.

[32] J. Hobbs, R. Nevatia, and B. Boles. An ontology for video event representation. In *Proc. IEEE Workshop Event Detection Recognit.*, page 119, 2004.

[33] S. Hongeng, R. Nevatia, and F. Brémond. Video-based event recognition: Activity representation and probabilistic recognition methods. *Comput. Vis. Image Understand.*, 96(2):129–162, 2004.

[34] M.-K. Hu. Visual pattern recognition by moment invariants. *IRE Trans. Inf. Theory*, 8(2):179–187, Feb. 1962.

[35] Y. Hu, X. Xie, W. Ma, L. Chia, and D. Rajan. Salient region detection using weighted feature maps based on the human visual attention model. In *Proc. IEEE PCM 2004*, pages 993–1000, 2004.

[36] S. S. Intille and A. F. Bobick. A framework for recognizing multi-agent action from visual evidence. In *Proc. Nat. Conf. Artif. Intell.*, pages 518–525, 1999.

[37] Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):852–872, Aug. 2000.

[38] S. W. Joo and R. Chellappa. Recognition of multi-object events using attribute grammars. In *Proc. Int. Conf. Image Process*, pages 2897–2900, 2006.

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

[40] A. Kale, A. Sundaresan, A. N. Rajagopalan, C. N. P., A. K. Roy-Chowdhury, V. Kruger, and R. Chellappa. Identification of humans using gait. *IEEE Trans. Image Process.*, 13(9):1163–1173, Sep. 2004.

[41] R. L. Kashyap. Bayesian comparison of different classes of dynamic models using empirical data. *IEEE Trans. Autom. Control*, AC-22(5):715–727, Oct. 1977.

[42] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 166–173, 2005.

[43] I. Laptev. On space-time interest points. *Int. J. Comput. Vis.*, 64(2-3):107–123, 2005.

[44] F. Liu and M. Gleicher. Region enhanced scale-invariant saliency detection. In *Multimedia and Expo, IEEE International Conference on*, pages 1477–1480, July 2006.

[45] M. C. Mazzaro, M. Sznaier, and O. Camps. A model (in)validation approach to gait classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(11):1820–1825, Nov. 2005.

[46] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(8):873–889, Aug. 2001.

[47] T. B. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Understand.*, 104(2):90–126, 2006.

[48] D. Moore and I. Essa. Recognizing multitasked activities from video using stochastic context-free grammar. In *Proc. 18th Nat. Conf. Artif. Intell.*, pages 770–776, 2002.

[49] T. Murata. Petri nets: Properties, analysis and applications. *Proc. IEEE*, 77(4):541–580, Apr. 1989.

[50] M. Negnevitsky. *Artificial Intelligence: A Guide to Intelligent Systems.* Adison-Wesley, 2nd edition, 2005.

[51] J. C. Niebles, H. Wang, and L. F. Fei. Unsupervised learning if human action categories using spatial-temporal words. In *Proc. British Mach. Vis. Conf.*, pages 1249–1258, 2006.

[52] B. North, A. Blake, M. Isard, and J. Rittscher. Learning and classification of complex dynamics. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(9):1016–1034, Sep. 2000.

[53] S. M. Oh, J. M. Rehg, T. R. Balch, and D. F. Data-driven mcmc for learning and inference in switching lienar dynamic systems. In *Proc. Nat. Conf. Artif. Intell.*, pages 944–949, 2005.

[54] S. Park and J. K. Aggarwal. Recognition of two-person interactions using a hierarchical bayesian network. *ACM J. Multimedia Syst.: Special Issue on Video Surveillance*, 10(2):164–179, 2004.

[55] V. Pavlovic, J. M. Rehg, and J. MacCormick. *Advances in Neural Information Processing Systems*, chapter Learning switching linear models of human motion, pages 981–987. MIT Press, Cambridge, MA, 2000.

[56] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference.* Morgan Kaufmann, San Francisco, CA, 1988.

[57] A. Pentland. Smart rooms, smart clothes. In *Proc. Int. Conf. Pattern Recognit.*, volume 2, pages 949–953, 1998.

[58] C. A. Petri. Communication with automata. DTIC Res. Rep. AD0630125, Defense Tech. Inf. Cntr., Fort Belvoir, VA, 1966.

[59] C. Pinhanez and A. Bobick. Human action detection using pnf propagation of temporal constraints. In *Proc. of Computer Vision and Pattern Recognition*, pages 898–904, 1998.

[60] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, Feb. 1989.

[61] N. Rota and M. Thonnat. Activity recognition from video sequences using declarative models. In *Proc. 14th Eur. Conf. Artif. Intell.*, pages 673–680, 2000.

[62] M. S. Ryoo and J. K. Aggarwal. Recognition of composite human activities through context-free grammar based representation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1709–1718, 2006.

[63] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer. The human id gait challenge problem: Data sets, performance, and analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(2):162–177, Feb. 2005.

[64] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[65] V. D. Shet, D. Harwood, and L. S. Davis. Vidmap: Video monitoring of activity with prolog. In *Proc. IEEE Conf. Adv. Video Signal Based Surveillance*, pages 224–229, 2005.

[66] Y. Shi, A. F. Bobick, and I. A. Essa. Learning temporal sequence model from partially labeled data. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1631–1638, 2006.

[67] Y. Shi, Y. Huang, D. Minen, A. Bobick, and I. Essa. Propagation networks for recognizing partially ordered sequential action. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, volume 2, pages 862–869, 2004.

[68] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(12):1371–1375, Dec. 1998.

[69] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time trakcing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):747–757, Aug. 2000.

[70] S. Tran and L. S. Davis. Visual event modeling and recognition using markov logic networks. In *Proc. IEEE Eur. Conf. Comput. Vis.*, Marseille, France, 2005.

[71] J. K. Tsotsos. Distributed saliency computations solve the feature binding problem. In L. Paletta, J. K. Tsotsos, E. Rome, and G. W. Humphreys, editors, *WAPCV2004: 2nd international workshop on attention and performance in computational vision*, 2004.

[72] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, November 2008.

[73] N. Vaswani, A. K. Roy-Chowdhury, and R. Chellappa. A continuous-state hmm for moving/deforming shapes with application to abnormal activity detection. *IEEE Trans. Image Process.*, 14(10):1603–1616, Oct. 2005.

[74] R. Vidal and P. Favaro. Dynamicboost: Boosting time series generated by dynamical systems. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1–6, 2007.

[75] P. A. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vis.*, 57(2):137–154, 2004.

[76] T. Wada and T. Matsuyama. Multiobject behavior recognition by event driven selective attention method. In *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, volume 22, 2000.

[77] A. D. Wilson and A. F. Bobick. Learning visual behavior for gesture analysis. In *Proc. Int. Symp. Comput. Vis.*, pages 229–234, 1995.

[78] A. Yilmaz and M. Shah. Actions sketch: A novel action representation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, volume 1, pages 984–989, 2005.

[79] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, volume 2, pages 123–130, 2001.

[80] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 819–826, 2004.