Bayesian Framework for Video Surveillance Application

Somboon Hongeng, Francois Brémond and Ramakant Nevatia
Institute for Robotics and Intelligent Systems
University of Southern California
Los Angeles, California 90089
{hongeng, nevatia}@iris.usc.edu, Francois.Bremond@sophia.inria.fr

Abstract

The goal of this paper is to describe and demonstrate the application of Bayesian networks in a generic automatic video surveillance system. Taking image features of tracked moving regions from an image sequence as input, mobile object properties are first computed and noise is suppressed by statistical methods. The probability that a scenario occurs is then computed from these mobile object properties through several layers of naive Bayesian classifiers (or a Bayesian network). Several issues and solutions regarding the efficiency of the Bayesian network are discussed. For example, the parameters of the networks, which represent rare activities (typical of video surveillance applications), can be learned from image sequences of similar scenarios which are more common. We demonstrate the effectiveness of our approach by training the networks with 600 image frames belonging to one domain of interest and applying them to image sequences in a different domain.

1. Introduction

Human activity recognition has received attention from computer vision researchers in recent years. This interest is motivated by a large number of real world applications including video surveillance in which a constant routine observation of the scenes by human operators is required. Automatic (or semi-automatic) activity recognition for video surveillance applications can relieve such tediousness or improve its efficiency. The task in automatic activity recognition for video surveillance applications is largely composed of 1) detecting and tracking the moving regions, and 2) recognizing the type of mobile objects and their activities. A critical point in step 2 is to map correctly the pixel level information (provided from detection and tracking of moving regions) to the high level symbolic description of activities. This is difficult because, firstly, it requires a powerful representation that supports the recognition of complex activities performed by humans. Secondly, the motion detection and tracking are often uncertain and incomplete, which requires recognition methods that can handle the probabilities accurately.

In a previous paper [4], we introduced a generic hierarchical representation and used it to represent and recognize several kinds of activities using a rule-based recognition method. In this rule-based method, rules are used to approximate the belief of the occurrence of activities. For example, they verify whether the properties of the mobile objects match their expected distributions (represented by a mean and a variance). This method often involves a careful hand-tuning of parameters such as threshold values.

The goal of this paper is to describe recognition methods that compute the probabilities of activities in a rigorous way using Bayesian networks. We focus on the computation of the recognition of scenarios which constitute the link between the low level image data and the high level complex activities. The paper is organized as follows. Related work is discussed in section 2. Section 3 briefly describes the hierarchical representation. Section 4 describes the activity recognition methods using Bayesian networks and related issues such as the learning of the network parameters and the limitation of Bayesian networks with regard to activity recognition. The results on real-world sequences of several domains of interest are shown in section 5.

2. Related Work

Current approaches to activity recognition are composed of defining activity models for specific types of activities that suit the goal in a particular domain and developing procedural recognition methods. In [7], simple periodic events (e.g. walking) are recognized by constructing dynamic models of periodic patterns of people's movements and is dependent on the accuracy of the tracking.

Inspired by a similar application to speech recognition, *Hidden Markov Model*(HMM) has also been applied to activity recognition [10, 3, 11]. For example, in [10], an HMM

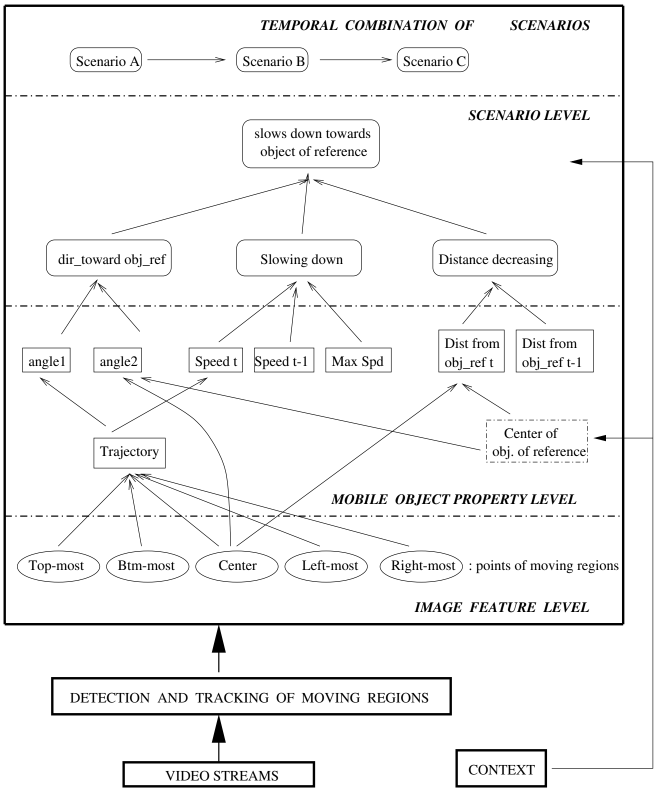


Figure 1. *Activity Recognition is composed of three levels of processing: image features of moving regions, mobile object properties and scenarios.*

is used as a representation of simple events which are recognized by computing the probability that the model produces the visual observation sequence. Even though HMMs are robust against various temporal segmentations of events, the structure and probability distributions are not transparent to human and need to be learned using iterative methods. For complex events such networks and their parameter space may become prohibitively large. Adapting an existing HMM to a different application (e.g. adding a new node) is also difficult.

Some other approaches use a Bayesian network as their representation [6, 2, 1, 8]. The use of Bayesian networks in these approaches differs in the way they are applied (e.g. what data is used as evidential input and how this data is computed, what are the structures of the networks, etc.). In [9], an action *sitting* is represented by the change in the position of a human head (tracked manually by humans). This network is simple and easy to train (fewer parameters) but it may not be able to discriminate similar actions such as *bending over*. In [8], Bayesian networks are used for recognizing several activities in a football match. However, no discussion about the learning issues of the network parameters is provided.

3. Activity Recognition Framework

To represent a wide variety of activities, we propose to use a hierarchical representation in which a Bayesian network can be integrated naturally. Figure 1 shows schematically how mobile objects and their behavior may be recognized from input visual data (image sequences) and available context. **Context** consists of associated information, other than the sensed data, that is useful for activity recognition such as a spatial map and prior activity expectation. From input image sequences, moving regions are detected and tracked, and several 2D (and, if available, 3D) **image features** (e.g. width, height, color histogram) are computed for these regions by lower level image processing routines. Mobile objects are then hypothesized as being composed of one or more tracked moving regions. *Several layers* of spatial and temporal **mobile object properties** (e.g. characteristics of a mobile object and short events such as *"entering the security area"*) are defined and computed over a few frames using specific methods (e.g. taking a ratio of width and height properties to compute the aspect ratio of the shape of mobile objects).

The distribution of mobile object property values are then used to compute the probability of **scenarios** that the mobile object may participate in. Scenarios correspond to long-term activities. Two types of scenarios are defined: single-state and multi-state. **Single-state scenarios** are defined by a *logical constraint* on a set of mobile object properties or sub-scenarios defined at lower layers. This constraint is verified from the evidence being observed at the current frame. For example, scenario *"mobile object* **A** *is slowing down towards mobile object* **B***"* can be verified at each frame from the fact that *"the moving direction of* **A** *is toward* **B***"*, *"***A** *is slowing down"* and *the distance between* **A** *and* **B** *is decreasing"*.

Multi-state scenarios are defined as a temporal combination of sub-scenarios to describe more complex activities such as a consecutive occurrence of scenarios (e.g. *"walks towards a person"*, then *"stops to talk"*, and then *"leaves"*). Multi-state scenarios can be represented by a chain of sub-scenario hierarchies and are recognized by verifying that all sub-scenarios are recognized corresponding to the combination type.

4 Scenario Recognition Methods

In our proposed representation, the recognition of complex activities relies on the recognition of single-state scenarios. Therefore, to improve the performance of the whole activity recognition systems, this paper is targeted at developing a recognition method for single-state scenarios that

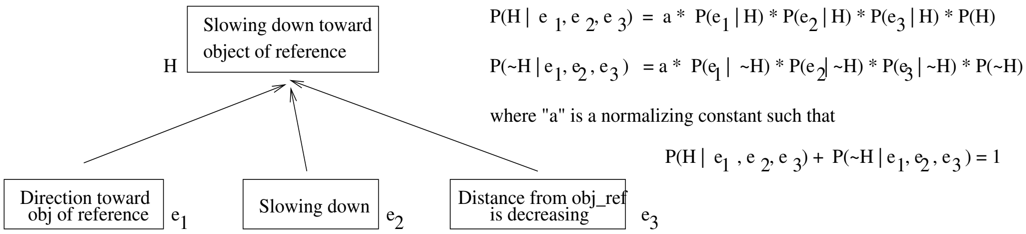


Figure 2. *A detailed illustration of a naive Bayesian classifier that is used to infer "slowing down toward object of reference" in figure 1. Given that e_1, e_2 and e_3 are conditionally independent given H, the belief is propagated from the sub-scenarios e_1, e_2 and e_3 to infer the probability distribution of H (i.e. $P(H|e_1, e_2, e_3)$) by applying Bayes' rule. e_1, e_2 and e_3 can also be a parent scenario of other naive Bayesian classifiers as shown in figure 1.*

computes the probabilities more accurately. The recognition methods of multi-state scenarios are described in [5].

Given a set of single-state scenario models defined by our proposed representation ($S = S_1, S_2, ..., S_N$) and a set of mobile object properties ($O = O_1, O_2, ..., O_k$) at time t, we determine the most likely scenario that may occur by computing $P(S_i|O)$ and select the one with the maximum likelihood. We consider a scenario as having a binary distribution (i.e. whether it occurs or does not occur). The distribution of $P(S_i|O)$ can be derived using Bayesian methods by combining the distribution of sub-scenario values from lower layers and propagating them towards the top layer. For example, in figure 1, the speed of the mobile object at time t is combined with the speed at time $t-1$ and the maximum speed to determine whether the mobile object is slowing down.

In Bayesian terms, the input entities are viewed as providing the evidence variables and the task is to compute the probability distribution of an output entity hypothesis. If the entities to be combined are statistically independent (given the scenario), a simple *naive* Bayes classifier can be used to compute the distribution of the combined result. When the entities are not conditionally independent, Bayesian networks offer an efficient representation that can represent dependence by decomposing the network into conditionally independent components. To effectively use Bayesian networks, we need the knowledge about the network structure (i.e. which entities are directly related or linked to which entities) and the conditional probabilities associated with the links.

In our case, the structure of the network is derived by heuristic means from the knowledge about the domain. For example, logical constraints of sub-scenarios that represent the recognition of a particular scenario indicate the direct causal link between them (i.e. the sub-scenarios are the consequences of the scenario). By defining each scenario such that its sub-scenarios are conditionally independent of each other given the scenario values, the hierarchy such as the one in figure 1 can be converted naturally into a Bayesian network which is composed of several layers of naive Bayesian classifiers. Each layer in the hierarchy can be viewed as having several naive Bayesian classifiers (one classifier per each scenario). Belief propagation (inference process) is performed in one direction from the bottom layer to the top layer.

Two layers of naive Bayesian classifiers are defined in figure 1. At the top layer, parent scenario *"slowing down toward object of reference"*, is linked to three child scenarios: *"direction is toward the object of reference"* (e_1), *"slowing down"*(e_2), and *"distance is decreasing"* (e_3). These child scenarios form another layer of three naive Bayesian classifiers (e.g. e_1 becomes a parent of *"$angle_1$"* and *"$angle_2$"*). The probability distribution over the parent scenario values in Bayesian classifiers ($P(H|e_1, e_2, e_3)$) is inferred from the distribution of child scenario values and the conditional probabilities of child scenarios given the values of the parent scenario (i.e. $P(e_1|H)$, $P(e_2|H)$, and $P(e_3|H)$ as shown in figure 2). By taking advantage of the fact that the nodes are transparent (e.g. we can observe whether the object is moving towards another object or whether it is slowing down), these conditional probabilities (i.e. $P(e_i|H)$ and $P(e_i|\neg H)$) can be learned from image sequences by making a histogram of observed values of the evidence variables, e_1, e_2, e_3, given the value of a given hypothesis, H.

Issues in Applying Bayesian Networks

Our framework allows a variety of visual evidence to be incorporated into the hierarchy as mobile object properties which are then combined in a Bayesian network. However, in real world applications, these properties are often noisy and incomplete due to shadows, occlusion, and the erratic motion patterns of the individual mobile objects. The characteristics of these effects normally have a wide variance and are difficult to estimate by learning from a limited number of examples. To minimize the errors caused by these artifacts, we provide filtering functions and mean functions that compute a mean value based on the multi-Gaussian distribution of the property values collected over time.

Another issue is concerned with the learning of the network parameters. In general, the number of training examples should be large enough so that the parameters are independent of the input image sequences. As the number of networks grows to accommodate the possible scenarios expected in a complex domain, the parameter learning task becomes tedious. However, when the relation between the parent node and the child node is obvious (for example, a strictly logical *"if and only if"* relation, or the range of the speed property of a vehicle), the conditional probabilities can be approximated. Therefore, in our applications, only a small number of network parameters are needed to be learned.

Another difficulty with regard to parameter learning is that the scenarios of interest in video surveillance applications are rare, which makes it impossible to collect enough examples. However, in our case, since Bayesian networks are applied at a symbolic level (i.e. scenario level), these parameters can be learned from a set of similar image sequences which are more common. For example, the network parameters for scenario *"passing through the checkpoint"* (to be detected in an army mission) can be learned from *"passing by a human"* image sequences taken from a parking lot.

The last issue is concerned with the application of Bayesian networks to recognize a temporal combination of durative scenarios (e.g. scenario **A** occurs before scenario **B**). Even though a Bayesian network can be used to combine the instantaneous probabilities of the sub-scenarios collected over a period of time, these values are normally not conditionally independent and the scenarios can have a wide variety of temporal segmentations which makes it difficult to estimate the necessary number of frames to be combined. Instead, we propose to recognize these types of scenarios at a higher level which is not described in this paper (see figure 1). By using Bayesian networks at a lower level to handle the uncertainty of mobile object properties accurately, the recognition of the activities defined at the higher level can be improved [5].

5. Results

We first show the recognition results of a mobile object from its spatial and temporal properties using a naive Bayesian classifier. Then, we show the effectiveness of the Bayesian networks (i.e. multiple layers of naive Bayesian classifiers) as applied to general activities by recognizing several scenarios in two different domains of interest: parking lot monitoring and checkpoint monitoring.

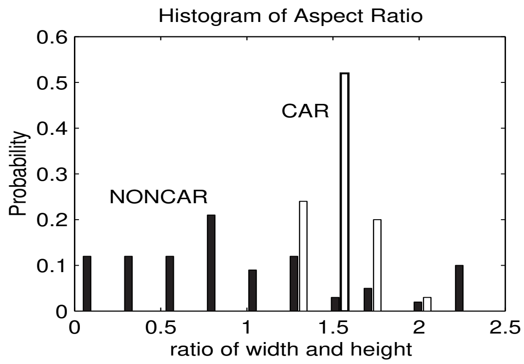
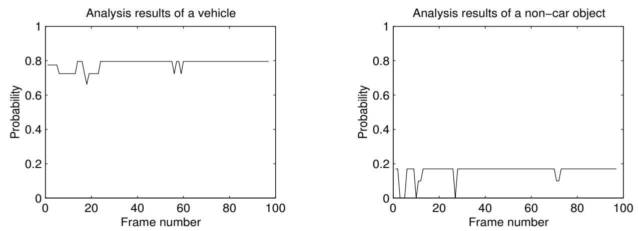


Figure 3. *Examples of the histogram of the aspect-ratio of mobile objects identified as a car (shown in white) and of other non-vehicle objects (shown in black).*



(a) Mobile object is a vehicle. (b) Mobile object is a human.

Figure 4. *Recognition results of "mobile object is a car" using a naive Bayesian classifier on two moving objects. (a) a moving object that represents a car is correctly recognized (P=0.8). (b) a moving object that represents a walking human is recognized as a car with very low probabilities.*

5.1 Recognition Results using a naive Bayesian classifier

To obtain the conditional probability distributions $P(e_i|H)$ and the prior probabilities $P(H)$ for each naive Bayesian classifier, we observe the occurrence of evidence e_i and hypothesis H and make a histogram of these quantities. For example, the hypothesis that the mobile object is a vehicle can be inferred from several spatial and temporal properties including the aspect ratio (between the width and the height) of the tracked moving regions. Figure 3 shows two histograms of the distribution of the aspect ratio obtained from 300 training image frames containing both detected vehicles and non-vehicle objects. These histograms represent the conditional probabilities ($P(ratio|car)$ and $P(ratio|non-car)$) associated with the links between scenario *"mobile object is a car"* and mobile object properties *"aspect ratio"*.

These parameters are then used to infer whether or not a mobile object is a car in different test video sequences. Figure 4 (a) shows the recognition results of a moving blob that represents a car. This blob is recognized as a car with a high probability through out the sequence. Figure 4 (b) shows the recognition results of a moving blob that represents a human in another sequence. This moving blob has a similar aspect ratio to that of a car. However, it is recognized as a car with a very low probability (i.e. 0.17) because the size of the blob does not match with the size of a car. These results indicate the effectiveness of the direct learning methods of network parameters that takes advantage of the transparency of the network structure. We believe that the recognition results from naive Bayesian classifiers (and also Bayesian networks) can be further improved by modeling the conditional probabilities (i.e. histograms) more accurately (e.g. increasing the quantization steps or using a parameterized model).

5.2 Discriminating Among Scenarios

We model activity **Contact1** as being composed of a consecutive occurrence of three single-state scenarios: *"going toward a meeting point"*, *"being close to a meeting point"* and *"turning around and leaving a meeting point"*, where *meeting point* is the location where another mobile object is located. These single-state scenarios (called scenarios 1, 2 and 3 accordingly) are constructed using Bayesian networks with a structure similar to the one shown in figure 1. The parameters of the networks are learned from sequences taken at a parking lot. The training data set is composed of 600 frames containing approximately an equal number of positive and negative examples. We analyze ten video streams that capture activities of interest in two different parking lot areas. The second parking lot has a different street structure from the first (i.e. the streets are of high curvature). Figure 5 shows a video sequence **A** (*"Contact1"*) taken at the second parking lot. By comparing the probabilities of all scenarios analyzed at each time frame, we notice that scenario 1 is recognized with the highest probability during frame 80 and 87, scenario 2 during frame 87 and 130, and scenario 3 during frame 130 and 150 accordingly. These results match with the human observation of the sequence.

5.3 Learning and Testing in Different Sequences

We also model a different activity called **Passing By** which is composed of three phases of single state scenarios: *"approaching an object of reference"*, *"being close to the object of reference"*, and *"going away"* (called scenario 1, 2 and 3 accordingly). Our goal is to train the Bayesian networks under a controlled environment to recognize a rare activity (i.e. in which case, only a limited number of examples are available). The parameters of the networks are learned in the same way as in the case of activity **Contact1** from several sequences taken from a parking lot. We then test the networks on the sequences taken from both the parking lot scene and from a rare checkpoint scene.

Figure 6 (a) shows sequence **B** and the recognition results of each scenario. Scenario 1 is recognized with the highest probability during frame 80 and 120, scenario 2 during frame 120 and 140, and scenario 3 during frame 140 and 180 accordingly. In comparison with the relatively smooth recognition results of sequence **A**, there are a few spikes in the recognition of sequence **B**. These spikes are due to the fact that the size of the moving objects is small and hence is sensitive to several artifacts including noise and shadows. Based on the temporal coherence property of data, these spikes may be smoothed out by combining the probabilities from different time frames. This can be done using an ad hoc method or a Bayesian network for an optimal result. We are currently exploring these alternatives.

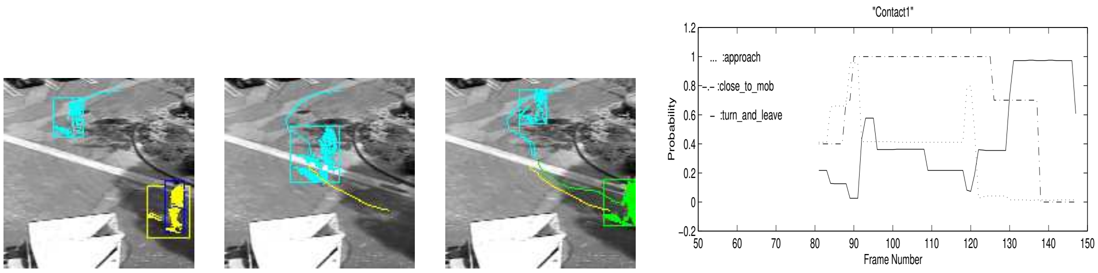
Figure 6 (b) shows the analysis results of test sequence **C** (taken at the checkpoint). The phase of each recognized scenario corresponds appropriately to the observation by humans. This experiment shows that the system can recognize scenarios even the parameters are learned from image sequences taken in different scenes.

6. Conclusion

We presented a new Bayesian framework to recognize various kinds of activities and differentiate between similar ones. This framework contains a hierarchical representation that describes scenarios from general properties of moving objects. These properties are then combined through Bayesian networks to infer the probability distributions of the scenarios. We have shown that the learning of the network parameters can be performed in a realistic way and that it does not depend on the training image sequences. Moreover, since Bayesian networks are applied at the scenario level (i.e. symbolic level), the parameters of the networks of uncommon scenarios can be learned from similar scenarios which are more common. The uncertainties of mobile object properties are computed accurately at the scenario level, which improves the ability of our approach to differentiate between similar activities. Our experiments indicate the validity of our approach. We plan to conduct more extensive tests to establish the limitation of our system.

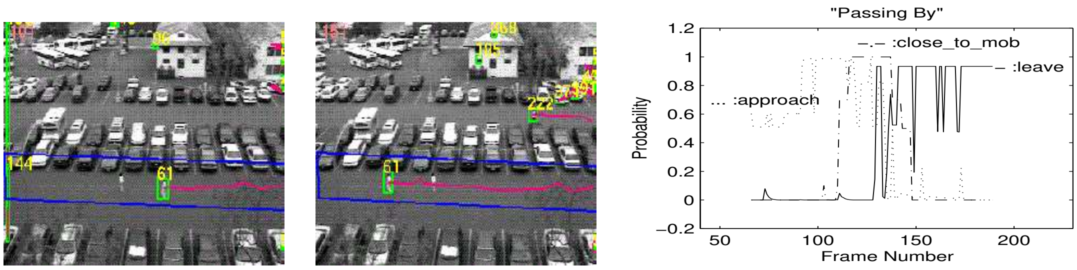
References

[1] P. Remagnino, T. Tan, K. Baker. Multi-agent visual surveillance of dynamic scenes. *Image and Vision Computing*, 16:529–532, 1998.



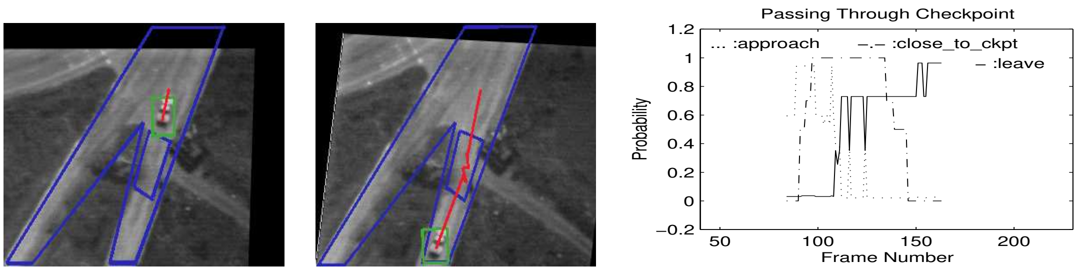
frame A.63 frame A.111 frame A.157 Recognition results

Figure 5. *Sequence***A** *shows object 1 (at the top) approaches object 2, makes contact, turns around and leaves. The most likely scenario can be inferred at each time frame by finding the one with the maximal probability.*



frame B.100 frame B.150 Result

(a) Input image sequence B and the recognition results of the components of scenarios "Passing By".



frame C.81 frame C.163 Result

(b) Applying the network to image sequence C of a different domain.

Figure 6. *(a) Input sequence* **B** *shows activity "Passing By" taken from a parking lot. Object 61 first approaches, then passes, and then leaves a meeting point where another mobile object is standing. The recognition of these sub-scenarios are shown accordingly on the right. (b) Composite scenarios of activity "car is passing through a checkpoint" are recognized by Bayesian networks whose parameters are learned from a similar scenario "passing by". The recognition results show that all three sub-scenarios are correctly recognized with high probabilities.*

[2] J. Binder, D. Koller, S. Russell, and K. Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29:213–244, 1997.

[3] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *IEEE Proceedings of Computer Vision and Pattern Recognition*, Puerto Rico, USA, 1997.

[4] F. Brémond and G. Medioni. Scenario recofnition in airborne video imagery. In *DARPA Image Understanding Workshop*, pages 211–216, 1998.

[5] S. Hongeng , F. Brémond and R. Nevatia. Representation and optimal recognition of human activities. In *IEEE Proceedings of Computer Vision and Pattern Recognition*, South Carolina, 2000.

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

[7] L. Davis, R. Chelappa, A. Rosenfeld, D. Harwood, I. Haritaoglu, and R. Cutler. Visual surveillance and monitoring. In *DARPA Image Understanding Workshop*, pages 73–76, 1998.

[8] S. Intille and A. Bobick. A framework for recognizing multi-agent action from visual evidence. In *IEEE Proceedings of the National Conference on Artificial Intelligence*, april 1999.

[9] A. Madabhushi and J. K. Aggarwal. A bayesian approach to human activity recognition. In *Second IEEE International Workshop on Visual Surveillance (CVPR Workshop*, pages 25–30, Fort Collins, Colorado, June 1999.

[10] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In *Proceedings of ISCV '95*, 1995.

[11] A. Wilson and A. Bobick. Recognition and interpretation of parametric gesture. In *IEEE Proceedings of the International Conference on Computer Vision*, Bombay, India, 1998.