Recurrent Bayesian Network for the Recognition of Human Behaviors from Video

Nicolas Moënne-Loccoz, François Brémond, and Monique Thonnat

INRIA Sophia Antipolis 2004, route des Lucioles, BP93 - 06902 Sophia Antipolis Cedex, France
{Nicolas.Moenne-Loccoz, Francois.Bremond, Monique.Thonnat}@sophia.inria.fr,
http://www-sop.inria.fr/orion/index.html

Abstract. We propose an original bayesian approach to recognize human behaviors from video streams. Mobile objects and their visual features are computed by a vision module. Then, using a Recurrent Bayesian Network, behaviors of the mobile objects are recognized through the temporal evolution of their visual features.

1 Introduction

Many works have used learning methods to recognize human activities from video streams in a cluttered, noisy and uncertain environment. None has solved the problem due to the temporal nature of human activities. We propose a learning method to recognize human behaviors from video streams in the context of metro station monitoring. Our approach is a particular form of a Dynamic Bayesian Network : a Recurrent Bayesian Network. A RBN models the temporal evolution of the visual features characterizing a human behavior and infers its occurence whatever its time-scale. In the second section we present the related work and show the need for a better temporal learning method. In the third section, we present the video interpretation system that computes the visual features used as input by the RBN, in the fourth section we formally present the RBN and in the fifth section we describe the results of our experiments.

2 Related Works

Learning methods have been used for the interpretation of video streams for the past ten years. There are two main approaches : probabilistic graph models and neural networks.

2.1 Probabilistic Graph Models

The probabilistic graph models allow to handle the uncertainty of the video processing task and to represent the prior knowledge of the application domain.

Bayesian Network BN are directed acyclic graphs. Each node represents a random variable and the links between the nodes represent a causality between random variables (e.g. A imply B) ; the links are associated to the conditional probabilities of that dependency. A BN is able to model the causalities between variables of a particular domain. Conditional probabilities are in general learned from a set of examples of the domain.

In [3], the team of H. Buxton uses BNs for the interpretation of video streams in a traffic monitoring system to recognize situation such as a traffic jam. BNs are used at two different levels : for the computation of simple but uncertain features of the scene and for the recognition of more complex behaviors. For example, a BN is used to infer the position of an object in the scene, using its orientation and its size. The orientation and the size of the object are inferred, in the same fashion, from visual features of the objects computed during the segmentation and the tracking process (object speed, object width ...).

The team of R. Nevatia [6] uses a naive bayesian classifier to recognize complex behaviors in the context of parking lot monitoring. For example the behavior *slowing down toward object of reference* is inferred from the events *moving toward object of reference*, *slowing down* and *distance from the object of reference is decreasing* by using prior probabilities computed during a learning phase.

BNs have the main advantage to use prior knowledge modeling the causalities between visual features as dependencies between random variables and to handle the uncertainty inherent to the video processing task.

Hidden Markov Models HMMs are a statistical tool for the processing of sequential data mainly used for the detection of pattern inside temporal sequences. They have been used successfully for speech recognition and recently for video interpretation. An HMM is a kind of finite state probabilistic automaton which transitions between states represent temporal causalities characterized by a probability distribution, generally learned from a set of examples of the domain. An HMM models a markovian temporal process, i.e. follows the Markov's hypothesis which states that a state depends only on the previous one.

The team of R. Nevatia [7] presents an approach using an HMM for the recognition of complex behaviors. A complex behavior is a behavior that can be represented by a sequence of simple behaviors. For example, the behavior *contact with an object of reference* can be represented by the sequence of behaviors *slowing down toward object of reference*, *contact with the object of reference* and *turn around and leave object of reference*. For that model, an HMM is constructed, which conditional probabilities are learned from a training set.

A.F. Bobick and Y.A. Ivanov [1] use a set of HMMs with a stochastic context free grammar to recognize simple gestures such as the movement of a hand drawing a square. Some HMMs are used to recognize elementary hand movements (*moving from bottom to top*,*moving from left to right*...) which sequence is syntactically analyzed by the grammar to recognize complex movements.

Team of A. Pentland [2] uses a particular form of HMM (*coupled HMM*) to recognize gestures of an asian gymnastic (*Tai'Chi*). Coupled HMMs are able to model interactions between processes such as the movements of the different hands of a subject. Hands gestures are independent but interact to form a *Tai'Chi* movement. But, coupling HMMs increases the learning complexity.

A. Galata, N. Johnson and D.Hogg [4] propose HMMs with variable length to model simple body exercises (*raise the left arm*, *flex the right arm*...) from the shape of the subject. The length of the HMM is chosen during the learning phase. The idea is, from a single state HMM, to evaluate the information gained by the use of a new state (increasing the length). Then, iterating the process, the best length is found, according to the evaluation of the information gain. Moreover, in order to recognize exercises of different level of abstraction, i.e. in different temporal scale, authors use a hierarchy of variable length HMMs.

Finally, an approach that uses hierarchical and unsupervised HMMs is presented by J. Hoey in [5]. The task of the presented system is to recognize facial expressions characterizing user emotions in order to interact with him. The originality of the approach is the unsupervise learning technique, that is learning from non-annotated examples.

HMMs model temporal dependencies of the phenomena. It is the learning method the most used for video interpretation, but HMMs are limited because of the Markovian hypothesis (most of human behaviors aren't markovian processes).

2.2 Neural Networks

Artificial neural networks are a set of simple information processing units (*artificial neurons*) that are organized in several layers and strongly interconnected. Each connection is characterized by a weight. Learning a processing task in a neural network is the task to find for a given network, the weights of the connections that model the best the processing task using a set of examples of that processing task.

H. Buxton [8] uses a Time Delay Radial Basis Functions Network. A TDRBF is a kind of neural network with an input layer, a hidden layer and an output layer. Neurons of the hidden layer are Gaussian radial functions. To provide a time delay, the network has as input a temporal window of the data which slides at each time. The TDRBF is used to recognize head movements from its rotation. The approach is interesting because it allows to learn a simple behavior on a time interval without knowing the model of that behavior. But the time interval is fixed and as the network structure, is chosen arbitrary. Moreover, the learning phase needs many examples : the experiments have used between 1000 and 5000 examples.

Only few works have used the possibilities of the neural networks for video interpretation. It seems to be an interesting solution because neural networks don't need prior knowledge but they need lots of examples.

3 Overview of a Video Interpretation System

We use a video interpretation system which is composed of two modules. The first module processes video streams : segmentation, classification and tracking of the mobile objects of the scene (essentially individuals or groups of individuals). The second module interprets the scene : it recognizes behaviors related to the mobile objects. Figure 1 presents an overview of the video interpretation system.

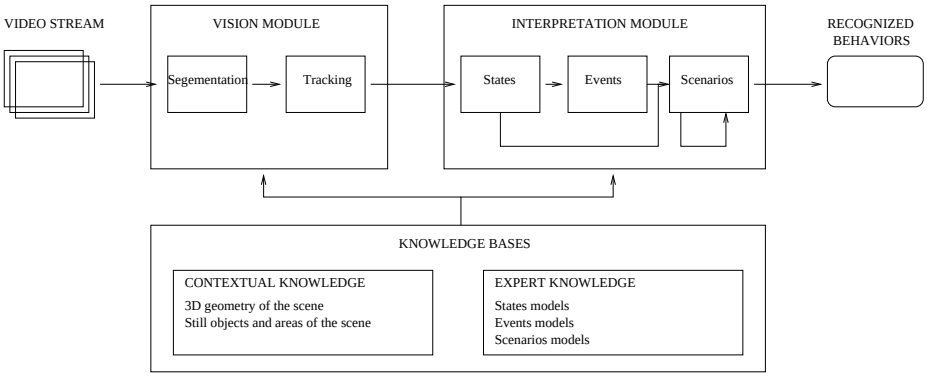

Fig. 1. Overview of the video interpretation system.

3.1 Vision

Processing of a video stream by the vision module provides a representation of the scene in terms of the mobile objects detected with a set of visual features characterizing these objects. More precisely, the task of this module is to :

- detect mobile objects of the scene,
- classify detected mobile objects (noise, individuals, groups of individuals),
- compute visual features of the detected mobile objects (size, direction, speed ...),
- track detected mobile objects as isolated individuals or globally as groups of individuals.

3.2 Interpretation

The task of the interpretation module is to recognize a set of behaviors that can occur inside the scene such as fighting or vandalism. As suggested by M. Thonnat and N. Rota [9], to ease the interpretation task, we introduce three entities which describe behaviors :

- State : a state is a mobile object property computed on a given time interval (*seating/standing*, *still/walking/running...*).
- Event : an event characterizes a change of state (*to sit down/to stand up*, *to stop/to begin running*). For example the event *to sit down* is the change of the state *standing* into the state *seating*.
- Scenario : a scenario is a combination of states, events and/or sub-scenarios (*graffiti on the wall*, *following someone*, *running toward the train...*).

In order to interpret the content of the scene, prior knowledge is provided to the interpretation system :

- Knowledge on the scene environment :
  - nature and position of the still objects of the scene (walls, benches, doors...),
  - semantic areas of the scene (platform, tracks...).
- Expert knowledge : states,events, scenarios models (*running toward the train = running + train present + trajectory is toward the train*).

From this prior knowledge and the representation of the scene provided by the vision module, the interpretation module use a bayesian approach to recognize hierarchically all occurrences of states, events and scenarios, i.e. all occurences of human behaviors.

4 Behavior recognition

4.1 Motivations

To recognize human behaviors from a set of visual features, we use a bayesian approach because of its ability to solve uncertainty, lack of knowledge and dependence on the scene environment. However, as seen in the section 2, the learning approaches for video interpretation are limited due to their inability to solve the problem of the temporal dimension. Effectively, there is no solution that handles the dependency on the time-scale of what has to be recognized. For example, it is necessary to be able to recognize a behavior lasting 10 seconds even if it usually lasts 20 seconds.

To overcome that limitation we propose an original form of a Dynamic Bayesian Network that is able to capture behavior whatever its time-scale.

4.2 Bayesian inference

A Recurrent Bayesian Network is a particular form of a Dynamic Bayesian Network dedicated to the recognition of behaviors. The behavior is inferred by the values of some visual features of the considered mobile object. The structure of a RBN takes as input the values of visual features during a fixed period of time and propagates the information of the previous periods of time through a recurrent dependency (previous occurrence of the behavior). Figure 2 shows the general structure of a RBN.

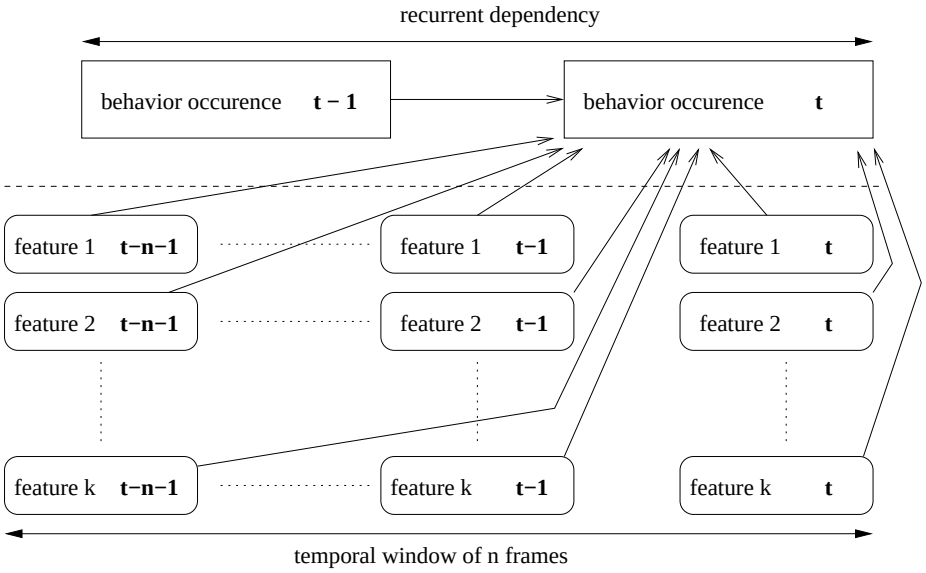

Fig. 2. *S*tructure of a Recurrent Bayesian Network.

The inference process in a RBN is the same as for a Bayesian Classifier. Formally, we have a behavior to recognize, B, and a set of dependencies O=$\{R, F1_t, F1_{t-1}, \ldots, F1_{t-n}, F2_t, F2_{t-1}, \ldots, F2_{t-n}, \ldots, Fk_t, Fk_{t-1}, \ldots, Fk_{t-n}\}$ where R is the recurrent dependency and Fi_t is dependency to the visual features i at time t. To simplify the formalism, we denote the dependencies by O=$\{O_1, O_2, \ldots, O_{nk+1}\}$ where n is the size of the temporal window and k the number of visual features used to infer the occurrence of the behavior. The idea is to compare the conditional probabilities $P(B|O)$ and $P(\bar{B}|O)$, and according to the maximum, decide weither B or $\bar{B}$ occures. From the Bayes theorem and if the O_i are independent conditionnaly to B (Bayes hypothesis) we obtain :

$$P(B \mid O) = \frac{P(O_1, O_2, ..., O_{nk+1} \mid B) * P(B)}{P(O_1, O_2, ..., O_{nk+1})} = \frac{\prod_i P(O_i \mid B) * P(B)}{P(O_1, O_2, ..., O_{nk+1})}$$

As we just need to compare $P(B|O)$ with$P(\bar{B}|O)$, the probability $\mathrm{P(O_1, O_2, \ldots, O_{nk+1})}$ is constant and has no influence on the comparison. And finally, in order to recognize B and $\bar{B}$ with the same probablility, we make the assumption that $P(B)$=$P(\bar{B})$. Finally we have :

$$\frac{P(B \mid O)}{P(\bar{B} \mid O)} = \frac{\prod_i P(O_i \mid B)}{\prod_i P(O_i \mid \bar{B})}$$

Probabilities $P(O_i|B)$ are computed during a training process, from a learning set $\{(b,o)_1, (b,o)_2, \ldots, (b,o)_m\}$ where each couple (b, o) represents a manually

annotated example, i.e. a frame characterized by the value of the behavior B and the value of all the dependencies O_i.

4.3 Network conception

Behavior reduction To model the network, the behavior as to be expressed as concrete as possible. For example, we are interessted to recognize a *violent behavior* involving a group of individuals in the context of metro station monitoring. A *violent behavior* is too abstract to be recognized. So we have to express it in term of a more concrete behavior. In our case, observing video streams that show groups involved in a *violent behavior*, we conclued that a group is saying having a *violent behavior* when it is globally agitated. Then we recognize a *violent behavior* as a high level of the *global agitation* of the group. Furthermore we observed that the *global agitation* of a group is the expression of two more simple behaviors : *internal agitation* and *external agitation*.

Features selection In order to recognize a human behavior such as a *violent behavior* we have to select a set of visual features from which one can infers the occurence of the behavior. Such a relation (*visual features* $\rightarrow$ *behavior*) appears to be not so obvious and a method is to use as many visual features as possible, knowing that selected visual features have to be independent conditionnaly to the behavior of interest. To increase the performance of the inference process, we consider only visual features relevant in regards to the conditionnal probabilites computed during the training process.

In our case, to recognize a *violent behavior* as a high level of *agitation*, we obtained the *recurrent bayesian network* of the figure 3.

The visual features used to infer the level of *global agitation* are :

- *Internal agitation* : the internal agitation is the visual agitation inside the group. From our experiments, we observed that internal agitation can be inferred by two visual features of the group :
  - *Evolution of the density* : the evolution of the ratio of the number of pixels that compose the group and the size of the group. It captures the agitation of the pixels inside the group.
  - *Movment of the center of gravity* : the movment of the center of gravity of all the blobs (set of pixels) of the group. It captures the agitation of the blobs relatively to each others.
- *External agitation* : the agitation of the individual relatively to the scene. External agitation is inferred by visual features that capture its movment in the scene :
  - *Evolution of the size (3d height * 3d width) of the group*
  - *Acceleration of the group*
  - *Evolution of the trajectory (orientation) of the group*

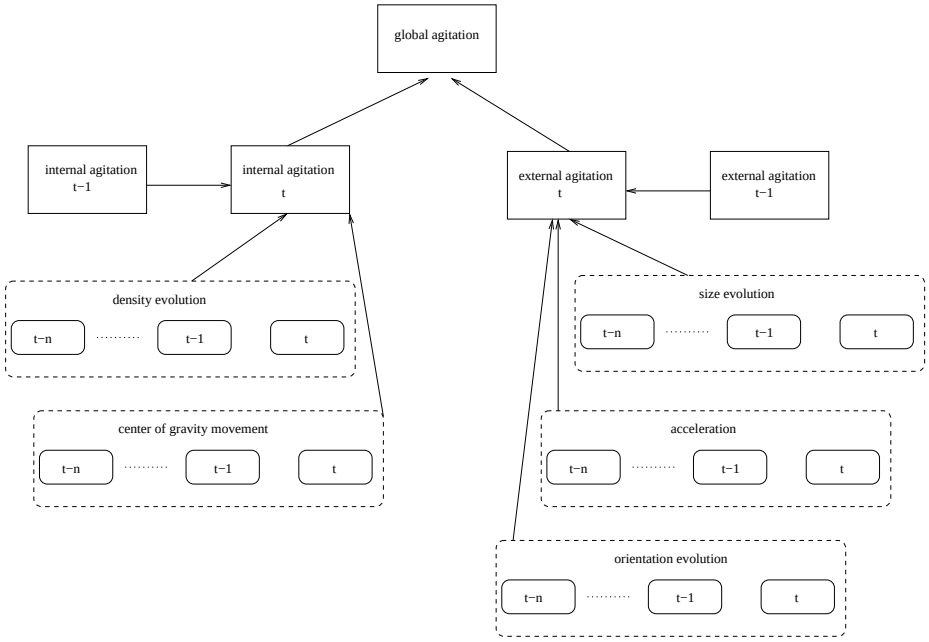

Fig. 3. Recurrent Bayesian Network for the recognition of the levels of the *global agitation* of a group.

The *internal agitation* and the *external agitation* are inferred using a RBN because they are time dependent behaviors. The *global agitation* is infered using a simple bayesian classifier because the time is already captured by the subsequent behaviors : *internal & external agitation*.

4.4 Learning

Annotation The process of annotating the video stream to compute the conditionnal probabilities presents some difficulties. The process is to decide for each frame if, given the information of the previous period of time, the behavior of interest is happening or not. A human operator has to take the decision on the basis of the observed period of time. Such an annotation is particularly subjective and repeating the process frequently gives different results, i.e. annotations.

Training Learning methods are limitated due to the lack of examples. To overcome this problem, we train the RBN with a set of nine video streams (a set of about 800 frames) and get the result for the tenth one (a set of 50-100 frames), iterating the process for each video stream.

5 Results

We validate the proposed formalism by analysing a set of video streams that show groups (of individuals) having either a *violent behavior* or not. These video streams are taken from the european ADVISOR project, in the context of monitoring a metro station.
As seen in section 4, a *violent behavior* is recognize as a high level of *global agitation* which is modeled by the RBN shown by the figure 3.

To train the network, we dispose of 10 video streams from the same site. 80% of the streams shows people fighting and the 20% others shows people not violent. The training set contains about 600 examples, i.e. frames that are annotated by a human operator according to the levels of *global, internal and external agitation* of the group present in the scene. Figure 4 shows the result for one of the stream with a group having a *violent behavior*, i.e. a high level of *global agitation*.

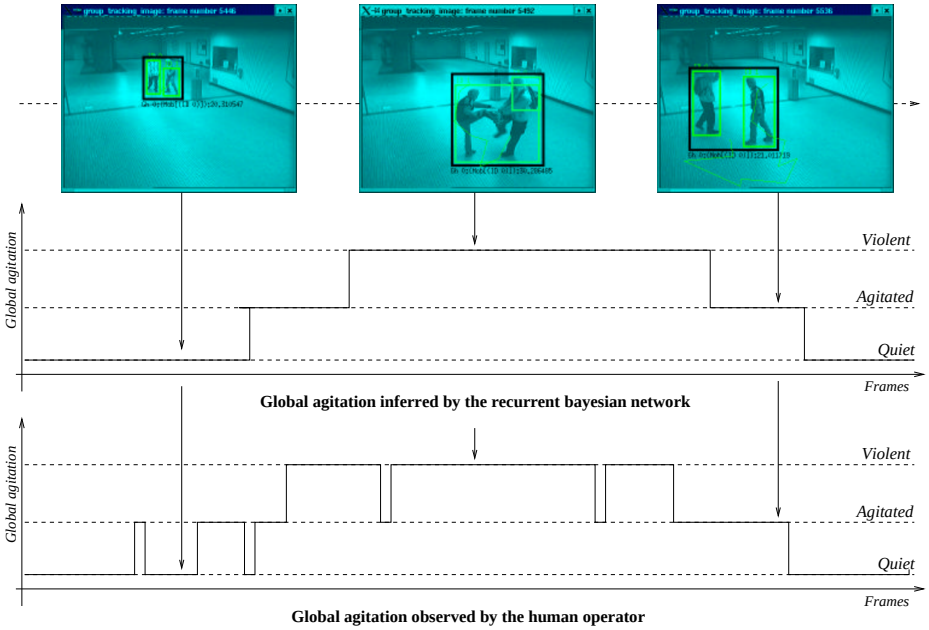

Fig. 4. Result for a sequence containing a *"violent behavior"*.

5.1 Performance

The temporal evolution of the *global agitation* of the group is correctly recognized. We observe two phenomena :

– a time delay : the recognition is done after a time delay, this is due to the time period in which we consider the visual features values.
– a time smoothing : ponctual changes are not recognized due to the recurrent dependency. This allows the recognition of the behavior to be made on the entire time period instead of for each frame.

Results are very encouraging : a *violent behavior* is recognized for each positive example and no *violent behavior* is detected for the two false examples.

5.2 Knowledge acquisition

Conditionnal probabilities computed during the training process give some knowledge about the domain of application. For example, we learned that the *external agitation* of a group is mainly inferred by the evolution of its size during the last 3 frames while its acceleration has only few influence and the evolution of its trajectory has a medium influence uniformly distributed on the temporal window.

6 Conclusion

We have proposed a Recurrent Bayesian Network to recognize human behaviors from video streams independently to their time-scale. We have validated the approach in the context of monitoring a metro station, for the recognition of *"Violent behaviors"*. Future works will validate the approach on several other behaviors in different contexts. Futhermore, as the learning process is tedious, we plan to use the RBN in an unsupervised mode.

References

1. A. F. Bobick and Y. A. Ivanov. Action recognition using probabilistic parsing. In *Computer Vision and Pattern Recognition*, 1998.
2. M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *Computer Vision and Pattern Recognition*, 1997.
3. H. Buxton and S. Gong. Advanced Visual Surveillance using Bayesian Networks. In *International Conference on Computer Vision*, 1995.
4. A. Galata, N. Johnson, and D. Hogg. Learning behavior models of human activities. In *British Machine Vision Conference*, 1999.
5. J. Hoey. Hierarchical unsupervised learning of facial expression categories. In *IEEE Workshop on Detection and Recognition of Events in Video*, 2001.
6. S. Hongeng, F. Brémond, and R. Nevatia. Bayesian framework for video surveillance application. In *International Conference on Pattern Recognition*, 2000.
7. S. Hongeng, F. Brémond, and R. Nevatia. Representation and optimal recognition of human activities. 2000.
8. A. J. Howell and H. Buxton. Recognizing simple behaviors using time-delay rbf network. 1997.
9. M. Thonnat and N. Rota. Image understanding for visual surveillance application. In *Workshop on Cooperative Distributed Vision*, 1999.