PARALLEL FIRST-ORDER MARKOV CHAIN FOR ON-LINE ANOMALY DETECTION IN TRAFFIC VIDEO SURVEILLANCE.

F.Archetti[1 2], C.E.Manfredotti[2], M.Matteucci[3], V.Messina[2], D.G.Sorrenti[2]

[1] Consorzio Milano Ricerche, Italy, via Cozzi 53, 20125 Milan archetti@milanoricerche.it
[2] Università Milano-Bicocca, Italy, via Bicocca degli Arcimboldi 8, 20121 Milan
{archetti, manfredotti, messina, sorrenti}@disco.unimib.it
[3] Politecnico di Milano, Italy, via Ponzio 34/5, 20133 Milan matteucc@elet.polimi.it

Keywords: Anomaly Detection, Traffic Video Surveillance, Markov Models, Behavior Modeling

Abstract

This paper focuses on on-line anomaly detection in video traffic surveillance systems. Markov Chain (MC) have been proposed already in computer and network intrusion detection. We applied them to the traffic domain and we propose to extend the classical MC (modeling all the behaviors in the scene) with an approach that evaluates in parallel a set of behavior specific MC. Such separate MCs are more discriminatory than a single MC for all the behaviors, allowing our approach to detect anomalies resulting from joining segments of normal behaviors. The learning of such models is done by using sequences of labeled normal behaviors and discretizing the image plane by using a simple grid. The approach has been validated on traffic surveillance videos, and experimental results show good performance both in terms of precision and recall.

1 Introduction

The last decade has seen a large increase in the use of video surveillance systems, installed to monitor the flow of pedestrians and vehicles, for security and data analysis. A lot of work has been done to develop autonomous surveillance systems which interpret the stream of images to recognize and classify activities from visual evidence.

Pioneering works of Nagel [9] and Neumann [11] have emphasized the need to deliver conceptual or symbolic descriptions of behavior from image sequences. Buxton and Gong [2] and Huang et al [7] follow this approach developing a visual surveillance system based on Bayesian belief networks that requires both metrical as well as topological information in order to interpret behavioral aspects of moving objects in a scene. More recently, the attention has been focused on probabilistic models as Dynamic Bayesian Network as shown by Gong and Xiang in [4] and Hongeng et al in [5] where such models are employed for activity recognition, while Hongeng and Nevatia [6] used semi Hidden Markov Models for event detection.

Anomaly detection techniques can follow two different approaches a *Signature* and a *Statistical* approach. Signature Anomaly Detection has an internal "list" of *anomalous* patterns; if an ongoing activity matches a pattern in the "list", an alarm is raised. Signature Anomaly Detection presents several disadvantages: the most important one is that, since the set of anomalous patterns is based on known anomalies, new one cannot be discovered. Statistical Anomaly Detection have been devised to address this shortcoming. The objective of Statistical Anomaly Detection is to establish profiles of *normal* activities: sequences of events that deviate from these profiles are considered anomalous and consequently an alarm is raised; an example of Statistical Anomaly Detection technique can be found in Vaswani et al [12].

Any statistical approach to Anomaly Detection follows this general strategy: using a set of normal patterns a statistical model is constructed; this statistical model is then used to construct a classifier that can discriminate between normal and anomalous tracks. The key point is that the statistical model should be an accurate predictor of normal behavior, so if an ongoing pattern is not accurately predicted by the model, it is likely to be anomalous. An interesting example of such approach can be found in Brand [1] where a Hidden Markov based Model (HMM) for video annotation is presented. In this paper traffic flow anomalies are detected on-line by looking for time windows in which the HMM assigns a very low Likelihood to the traffic flow vectors associated to the incoming images. Despite their high representation and generalization power, the performance of HMM based models is strongly dependent on their topology structure. Topology structure learning is a NP-hard problem and usually it is selected manually through experimental trials. Another problem that makes HMM very difficult to be applied operatively is parameter initialization, which strongly affect the subsequent learning phase.

The models above can be used to detect occurrence of suspicious events when the possible events are known. Our interest concerns something slightly different w.r.t. representing any possible configuration of moving objects. We are interested in the more limited case of detection of anomalous single driver behaviors, i.e. behaviors that does

not involve more than one vehicle. Therefore we aim at identifying, tracking and monitoring, independently, each vehicle moving in the scene. We, therefore, need a classifier, to discriminate if a track, i.e. a vehicle trajectory in the scene, corresponds to a "normal" or to an "anomalous" driver behavior.

In our application for traffic video surveillance we deal with on-line single driver behavior identification and anomaly detection. This activity is part of a more complex traffic monitoring system based on a layered architecture composed of modules extracting features of increasing level of abstraction from video sequences. From the lower to the higher level of abstraction we have: a *motion detection* module in charge of detecting moving objects from the frames of the video sequence, a *tracking* module in charge of following targets moving in the scene, and the *classification* module in charge of distinguishing between normal and anomalous tracks/behaviors. This final module consists in a *track classifier* based on Markov Chain (MC, here after) that detects anomalies that have to be brought to human attention.

Moreover, in the design of such surveillance system, we are interested in finding the simplest possible model for an on-line processing. According to this philosophy, the simple MC-based models, without hidden states, have been proved to be competitive in several applications regarding intrusion detection in computer and network systems. In particular, Nassehi [10] and Jha et al [8] adopted first and third order MC, respectively, to detect anomalies in a telecommunication system.

We applied MCs to the traffic domain and soon discovered the need for an improvement, which is proposed in this paper. In the traffic domain, anomalous patterns are frequently a composition of sub-parts of normal behaviors (see Fig. 4, left and right). These patterns are difficult to detect and could even be the only anomalies to be detected. We can call the problem of detecting such patterns "the normal sub-parts composition problem". Our approach handles multiple hypotheses in order to detect also such patterns. This is obtained by the use of different simple Markov Chain models, one for each correct behavior, trained with set of sequences from object's normal behaviors. Each of these probabilistic models can be used as a sequence classifier; the Likelihood of the observed sequence, given by the Markov Chain, can be used to discriminate between multiple hypothesis on the objects behavior and thus among normal and anomalous sequences of activities.

One could think that the normal sub-parts composition problem could be dealt with an increase in the order of the single MC used. This is partially true, with a provision. The order should be adapted to the specific velocity of the tracked vehicles and to the geometry of the observed scene. One could even try overkilling, i.e. guessing an upperbound on all possible intersections of normal behaviors in all possible scenes. In our opinion this severely affects the actual usefulness of solving the normal sub-parts composition problem by increasing the order of a single MC.

The next sections are organized as follows: in Section 2 basic Markov Chain definitions are given, along with the description of the modeling, learning and training phases. Section 3 describes the parallel Markov Chain model while Section 4 contains some experimental results which demonstrate the validity of the proposed approach. Conclusions are derived in Section 5.

2 Sequence Modeling with Markov Chains

A Markov Chain (MC) is formally defined as a tuple $\langle \mathcal{S}, \mathcal{T} \rangle$; where $\mathcal{S} = \{s_0, s_1, \cdots, s_M\}$, is a finite set of states and $\mathcal{T} : \mathcal{S} \times \mathcal{S} \rightarrow [0, 1]$, is a transition model, that is the probability for the system to get in a state s' at time $t+1$ being that it was in state s at time t:

$$\mathcal{T}(s, s') = P(s_{t+1} = s' | s_t = s) \tag{1}$$

with $\sum_{s' \in \mathcal{S}} P(s_{t+1} = s' | s_t = s) = 1, \forall s \in \mathcal{S}$. This simple transition model $\mathcal{T}$ assumes the whole system representation at time $t+1$ (s_{t+1}) depending only on the system description at the previous time step (s_t), and not on the whole previous history of the system ($s_t, ..., s_0$). In other words, the environment state s_t at time t is a sufficient statistic for the history of the system; this is the usual *Markov Property*:

$$P(s_{t+1} = s \mid s_t, ..., s_0) = P(s_{t+1} = s \mid s_t) \tag{2}$$

In real-world applications, even when the state is non-Markov, it is often still appropriate to consider the system behavior as approximated of a Markov process. Equation (2) defines a first order Markov process, we can consider an extension to the n^{th} order of the MC, where the Markov Property becomes:

$$P(s_{t+1} = s \mid s_t, ..., s_0) = P(s_{t+1} = s \mid s_t, ..., s_{t-n}) \tag{3}$$

meaning that system state at time $t+1$ (s_{t+1}) depends only on previous n system configurations ($s_t, ..., s_{t-n}$).

Given a sequence $\mathbf{s} = (s_0, s_1, \ldots, s_K)$ of states, we can compute its probability according to the MC model using the Chain rule and the Markov Property (Equation 2):

$$\begin{aligned} P(s_0, \ldots, s_K) &= P(s_K \mid s_{K-1}, ..., s_1) \cdot P(s_{K-1}, ..., s_1) \\ &= P(s_K \mid s_{K-1}) \cdot \ldots \cdot P(s_1 \mid s_0) P(s_0) \\ &= P(s_0) \cdot \prod_{k=1}^{K} P(s_k \mid s_{k-1}) \end{aligned} \tag{4}$$

where $P(s_0)$ is the initial state probability distribution.

A MC can be also described by an oriented graph $G(V, E)$, where V represents a set of nodes associated with the states of

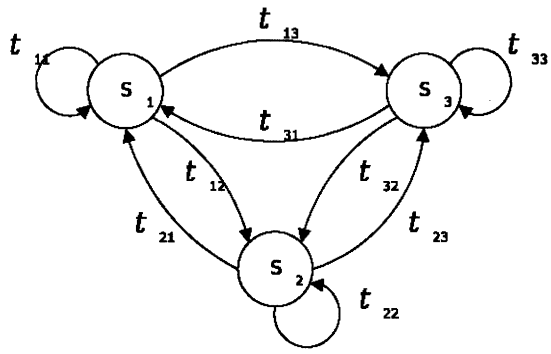

Figure 1: A graphical representation of a MC

the MC and E is the set of arcs (s_i,s_j) with s_i and $s_j \in V$. Each arc (s_i,s_j) is weighted with the probability of transition from state s_i to state s_j as shown in Fig. 1. The probability of all edges outgoing from a node sums up to one. For a through introduction to MC refer, e.g., to Durrett [3]

It is possible to learn the MC model from a data set of sequences. The probability transition matrix T can be defined as an M dimensional square matrix, being M the cardinality of S, whose element $t_{i,j}$ represents the transition probability from state s_i to state s_j. Recalling that the probability of all the transitions outgoing from a state must sum up to 1, we can estimate such multinomial distribution as:

$$t_{i,j} = \frac{n_{i,j}}{\sum_{k=1}^{M}(n_{i,k})} \tag{5}$$

where $n_{i,j}$ is the number of times the transition from s_i to s_j occurred in the training set.

Equation (5) implies $t_{i,j} = 0$ for all unseen transitions. To overcome this issue we introduce a small bias to this estimate by assigning a smoothing prior transition probability p among all possible transitions as follows:

$$t_{i,j} = \frac{n_{i,j} + p}{\sum_{k=1}^{M}(n_{i,k} + p)} \tag{6}$$

The effect of this *prior* assumption vanishes with the increase of observed transitions. Whenever some knowledge about the scene in available we can force to 0 the probability of transitions that are known to be impossible.

In order to estimate the initial state probability distribution, we count the times s_i occurred as initial state for the tracks in the training set (n_i), and we add a uniform smoothing prior q to this estimate:

$$p_i = \frac{n_i + q}{\sum_{j=1}^{M}(n_j + q)} \tag{7}$$

Given a sequence of states s, $(s_0, ..., s_k, ...s_K)$ we can compute its Likelihood $L(\mathrm{s})$, given the model, as the product of the probability of all the transitions multiplied by the probability of the initial state:

$$L(\mathrm{s}) = p_0 \prod_{k=0}^{K-1} t_{k,(k+1)} \tag{8}$$

For computational reasons it is common practice to use the Loglikelihood $(\mathcal{L}(\mathrm{s}))$:

$$\begin{aligned} \mathcal{L}(\mathrm{s}) &= \log p_0 \prod_{k=0}^{K-1} t_{k,(k+1)} \\ &= \log p_0 + \sum_{k=0}^{K-1} \log t_{k,(k+1)} \end{aligned} \tag{9}$$

$L(\mathrm{s})$ (or, equivalently, $\mathcal{L}(\mathrm{s})$) represents a measure of how well the model explains the given sequence of states. If $L(\mathrm{s})$ is lower than a given threshold τ we can infer that the model does not recognize s as a normal sequence. Therefore, given a threshold τ, we ca write the classifier D as

$$D(\mathrm{s}) = \begin{cases} 1 & \text{if } L(\mathrm{s}) < \tau \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

if the classifier is set to 1 an alarm will be raised.

This approach can recognize an anomalous sequence only when complete. In order to classify sequences on-line, we need to normalize the Log-Likelihood w.r.t. the number of transitions occurred in the ongoing track using the *Average Log-Likelihood* $(\overline{\mathcal{L}})$ of the sequence:

$$\overline{\mathcal{L}}(\mathrm{s}) = \frac{1}{n}\left(\log p_0 + \sum_{k=0}^{n-1} \log t_{k,(k+1)}\right) \tag{11}$$

where n represents the number of transitions occurred in the ongoing track.

3 The Parallel Markov Chain Classification Model

Working with a sequence of images leads us to deal with sequences of $2D$ object positions on the image plane; these sequences of positions, called tracks, come as output of the tracking step. In order to map the tracks on the discrete MC classification model we divide the image into a grid. A state of the MC is thus represented by a *cell* of the grid (see Fig. 2 or Fig. 5). Therefore, given a set of states S, a *track* s is an ordered sequence of states $\{s_1, ...s_k, ...s_K\} \in S$.

We name the cells of the grid with a label, using integer numbers (it is just a convention) and map the 2D image positions (x, y) into grid cells (*label/integer number*) by the relation:

$$label = \lfloor x/H \rfloor * N + \lfloor y/W \rfloor \tag{12}$$

where W and H are the horizontal and vertical dimension of the cells and N is the number of cells in the horizontal size:

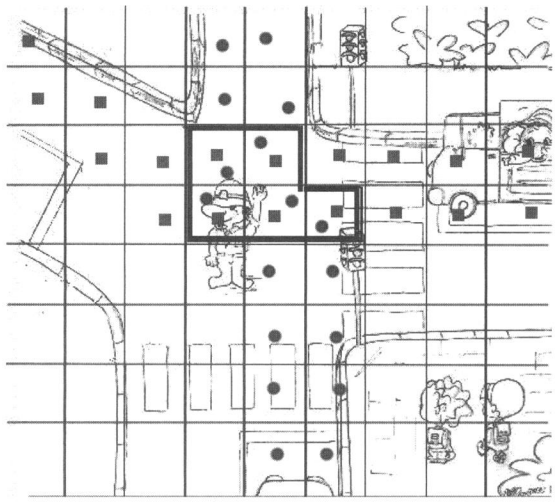

Figure 2: In a cross roads different behaviors can be considered as normal. The red squared represent the possible states for the normal behavior (see Fig. 3 (right)); the blue points represent the possible states for the another normal behavior (see Fig. 3 (left)). The states labeled by both are states belonging to the intersection of the two behaviors.

given the horizontal dimension of the image, X, $N = \frac{X}{W}$. The level of discretization depends on the precision one wants to obtain. From now on with "track" we mean a sequence of 2D image positions (x, y) mapped into labels, corresponding to a sequence of states of arbitrary length K.

Since we are interested in detecting anomalies of single vehicle behavior, the plain MC introduced in the previous sections is not sufficient, as we will make evident. If we consider a scene like, for example, the one in Fig. 2 we can cluster similar tracks into groups, representing different normal behaviors.

definition 1 ***Behavior***

We call Bh_i the set of all tracks corresponding to the normal behavior i. Being a track a set of states, we can also define a set B_i as the set of states belonging to the tracks of Bh_i, i.e. if $\mathbf{s}_j^{Bh_i}$ is the generic j-th track in Bh_i, $B_i = \bigcup\{s_k : s_k \in \mathbf{s}_j^{Bh_i}\}$.

The approach presented in the previous sections (the MC-based system) will find a false normal track each time an anomalous track is a combination of normal sub-tracks, no matter the order of the MC. Consider, for instance, two normal tracks (Fig. 3 left and right) and suppose that the van cannot turn on its right or that the car cannot turn left (see Fig. 4, left and right). An anomaly detector system, while analyzing a track like these, must return an alarm. However, the probability of this tracks computed as in Equation (11), is larger than τ for a MC-based system, because the track is composed by "sub-sequences of normal tracks" and (see Equation (10)) the alarm will not be raised. Such situation is very frequent in real world.

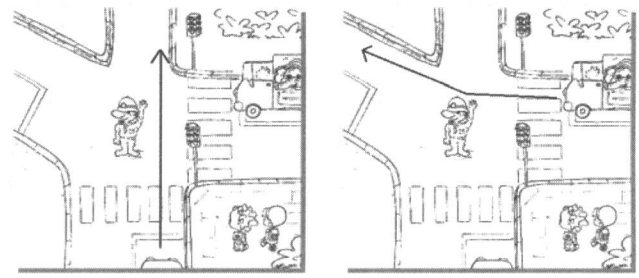

Figure 3: Two examples of normal behaviors

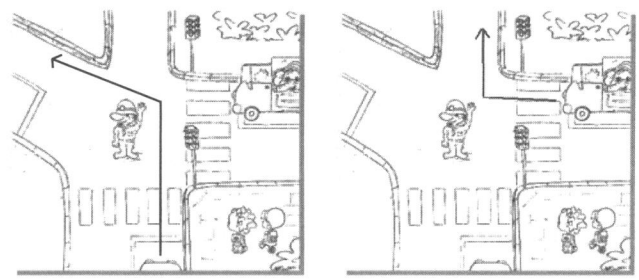

Figure 4: Two examples of anomalous behaviors

More formally:

proposition 1 *Given a track* $\mathbf{s}$*, if* $\exists\, Bh_i \subseteq \mathcal{S}$ *such that* $\forall s_k \in \mathbf{s}$*,* $s_k \in B_i$*,* $\mathbf{s}$ *is* normal.

proposition 2 *Consider two Bh, $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A}$ and $\mathcal{B} \subset \mathcal{S}$ with $\mathcal{A} \cap \mathcal{B} \neq 0$, $\mathcal{A} \nsubseteq \mathcal{B}$ and $\mathcal{B} \nsubseteq \mathcal{A}$. A track $\mathbf{r}$ $(r_0, ..., r_a, r_l, ..., r_m, r_b, ..., r_K)$ such that:*

- $(r_0, ..., r_a) \in \mathcal{A} \setminus \mathcal{B}$
- $(r_b, ..., r_K) \in \mathcal{B} \setminus \mathcal{A}$ *and*
- *the sub-sequence* $(r_l, ..., r_m) \in \mathcal{A} \cap \mathcal{B}$

is anomalous.

proposition 3 *Given an anomalous track $\mathbf{r}$ with elements in $\mathcal{S}$, an approach that models the normal behaviors with a unique MC of the n^{th} order will not detect $\mathbf{r}$ as anomalous if (and only if) the sub-sequence $(r_l, ...r_m)$ of proposition 2 is a sequence of at least n elements.*

To overcome such failure problem, we adopt a multiple hypothesis approach: we model each normal behavior in the scene with a MC, and we train a probability transition matrix for each behavior/MC. We evaluate a track on each model in parallel and, if the probability of the track is less than τ for each model, the system would return an alarm. Given a track $\mathbf{s}$, called $\overline{\mathcal{L}}_i(\mathbf{s})$ the Average Log-likelihood (Equation (11)) for

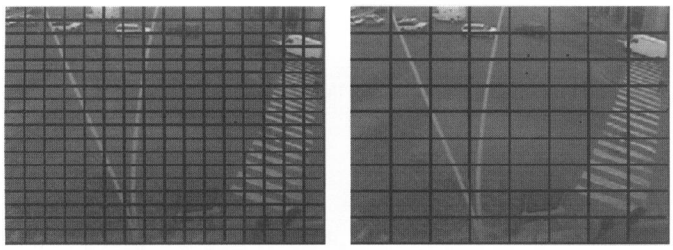

Figure 5: The cross roads, fine-grained one (left) and coarse-grained one (right)

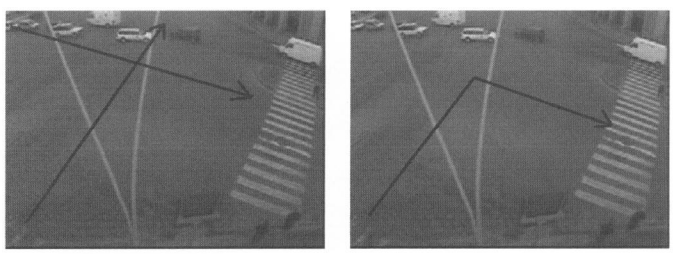

Figure 6: The cross roads, two examples of behaviors, two normal one (left) and an anomalous one (right)

the MC model learnt from the behavior Bh_i, the alarm will be raised for each $\mathbf{s}$ that satisfies: $\overline{\mathcal{L}}_i(\mathbf{s}) \leqslant \tau,\ \forall i$.

If we model each allowed behavior (Fig. 3) with its own MC, the tracks in Fig. 4 will be considered anomalous.

The number of MCs used to model normal tracks depends on the number of normal behaviors allowed in the scene.

4 Experimental Results

To validate our approach we conducted some experiments on many cross-roads video sequence. A cross-roads is the typical example for which an approach that uses a unique MC might fail showing the advantages of our approach. In this particular experiment we are interested in detecting cars coming from south and turning right (see Fig. 6 right).

The video sequences analyzed are composed by images of 288×384 pixel. We report here an example. We conducted a first experiment using a grid of 16×24 cells and a second experiment using a grid of 32×48 cells comparing the two approaches: a first order MC model modeling all the normal tracks with a unique MC and our method.

We considered 5 minutes of video footage and manually selected 3 minutes of video sequences containing only normal behaviors (we selected the video sub-sequences containing only normal behaviors). This normal behaviors have been used to train the models to be tested on the entire video sequence.

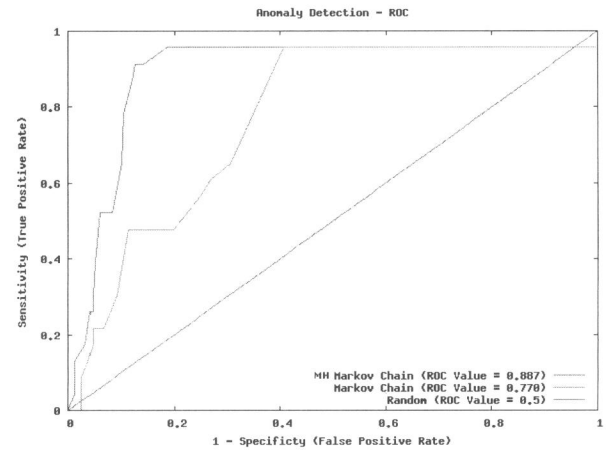

Figure 7: The cross roads. ROC curves of the two approaches on the fine-grained grid. The ROC value of an optimal method is equal to 1: the ROC value of an approach using two MCs is larger than the one of an approach that uses only a MC.

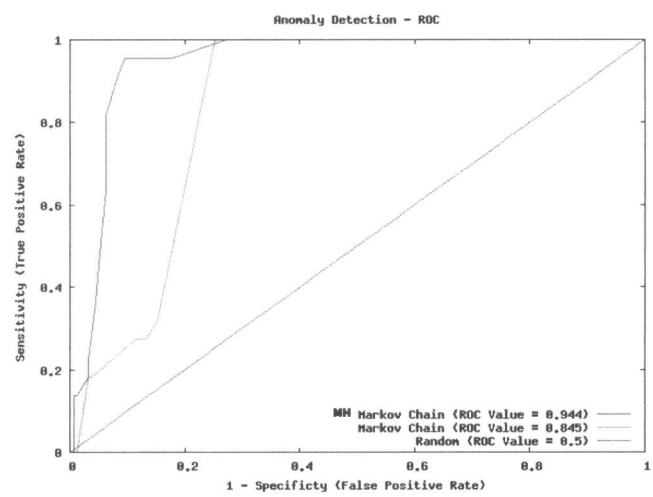

Figure 8: The cross roads. ROC curves of the two approaches on the coarse-grained grid.

The full sequence contains 90 tracks; of these 45 are vertical, 25 are horizontal from the left and the remaining 10 are considered as anomalous. An anomalous track is a track of a car that turns right coming from the bottom of the image: the first half of these tracks is well modeled by a model built by the set of vertical tracks, the second half by the set of horizontal tracks.

The first 3 minutes of the video show 54 tracks, 36 are tracks belonging to cars driving from the south and 18 in horizontal direction. We used all these tracks to instruct the unique MC and the two sets separately to instruct two different MCs.

We set the prior $p = 1$ and computed the matrix T to make all the transitions be possible. The two methods have been compared by using a ROC curve on both the fine-grained and the coarse-grained grid.

The Receiver Operator Characteristic curve (ROC curve) is a plot of the true positive rate against the false positive rate for the different possible values of threshold of the test. The closer the curve follows the left-hand border and than the top borders of the ROC space, the more accurate the test is. The area under the curve is a measure of the accuracy of the approach.

The ROC curves in Fig. 7 and in Fig. 8 show the performance of the two methods on the variation of the threshold τ on the two grids. In this scenario it is possible to notice how an approach modeling normal behaviors with different MCs outperforms the approach that uses a single MC.

5 Conclusions and future work

In this paper an approach to detect anomalous behaviors in traffic video surveillance has been presented. This approach extends the classical MC model taking multiple hypothesis about the observed behaviors.

The classical first order MC (modeling all the behaviors in the scene) has been extended by the evaluation of a parallel set of behavior specific MC models.

This approach turns out to be more accurate than an approach based on a single first-order MC trained on all behaviors since the latter fails in detecting anomalous behavior that are composed of normal behavior parts. The former, instead, is able to discriminate also such anomalous behaviors.

The approach has been validated on traffic surveillance videos, and experimental results show good performance both in terms of precision and recall. Learning of models has been done by using sequences of labeled normal behaviors and results have been shown on 288×384 images discretized by using different grids on the image plane. From the comparison of the ROC curves it is possible to notice that our multiple hypothesis approach outperforms classical MC and that different (reasonable) grids used to discretize the tracks have little effect on the final performance of the anomaly detector.

Presently we are working on the use of an adaptive grid to discretize the image plane according to track distribution (on the training set). This would improve model performance by taking into account perspective and specific scenario characteristics. This kind of grid could also be used to model the speed of the tracked vehicles.

The main disadvantage of our method is the need for a hand-labeled data set of "normal" behavior tracks. In a complex traffic situation, the task of labeling such behaviors could become too demanding. To ease this activity an unsupervised approach (i.e., a clustering algorithm) could be used.

Regarding the model we think that our approach leads to performance fully comparable to HMM but presents complete understanding of the internal states, we are actually performing this comparison on the videos presented in this paper.

References

[1] Matthew Brand. The "inverse hollywood problem": From video to scripts and storyboards via causal analysis. In *AAAI/IAAI*, pages 132–137, 1997.

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

[3] R. Durrett. *Probability: theory and Examples*. Duxbury Press, 3rd edition, 2004.

[4] Shaogang Gong and Tao Xiang. Recognition of group activities using dynamic probabilistic networks. In *ICCV*, pages 742–749, 2003.

[5] Somboon Hongeng, François Brémond, and Ramakant Nevatia. Representation and optimal recognition of human activities. In *CVPR*, pages 1818–1825, 2000.

[6] Somboon Hongeng and Ramakant Nevatia. Large-scale event detection using semi-hidden markov models. In *ICCV*, pages 1455–1462, 2003.

[7] Timothy Huang, Daphne Koller, Jitendra Malik, Gary H. Ogasawara, B. Rao, Stuart J. Russell, and Joseph Weber. Automatic symbolic traffic scene analysis using belief networks. In *AAAI*, pages 966–972, 1994.

[8] Somesh Jha, Kymie M. C. Tan, and Roy A. Maxion. Markov chains, classifiers, and intrusion detection. In *CSFW*, pages 206–219, 2001.

[9] H. H. Nagel. From image sequences towards conceptual descriptions. *Image and Vision Computing*, 6:59–74, 1988.

[10] M. Nassehi. Anomaly detection for markov models. Technical report, 1998.

[11] B. Neumann. Natural language description of time varying scenes. *Semantic Structures*, (167-206), 1989.

[12] Namrata Vaswani, Amit K. Roy Chowdhury, and Rama Chellappa. Activity recognition using the dynamics of the configuration of interacting objects. In *CVPR (2)*, pages 633–642, 2003.