

# A Step to Cognitive Vision Systems for Common Videosurveillance

Xavier Desurmont, Jerome Messen, Christophe Parisot, Jean-François Delaigle  
{desurmont, messen, parisot, delaigle}@multitel.be, Multitel A.S.B.L, Mons, Belgium.

## ABSTRACT

Automated videosurveillance for homeland security is a very hot topic today. Commercial systems are emerging but the configuration setup time and installation cost are still major issues. These systems usually detect unexpected or unauthorised events in a visual scene (e.g. unattended object, human motion) and then trigger alarms or record forensic video data and metadata in tamperproof databases. Currently, the configuration phase requires the intervention of experts for encoding precisely how the system has to work. In our approach we propose new functional capabilities for intelligent video-surveillance systems to learn events of interest from a base of representative examples. Such cognitive vision systems are based on basic capabilities like object detection, characterisation and tracking for low-level (image processing) and high-level (semantic) event recognition. We show how this semantic domain is useful to generalise the recognition. With this goal in mind, a performance evaluation of the system is performed and validated on test sequences including CAVIAR.

## 1. INTRODUCTION

Homeland security using automated systems is what users want. Videosurveillance is one market where automatic systems are becoming a reality. Examples of challenging applications [1] are monitoring metro stations [2] or detecting highways traffic jams as well as detecting loitering. Actually, an important disadvantage of such a system is the cost of the configuration step because it requires the intervention of experts for encoding precisely how the system has to work. In our approach we propose new functional capabilities for intelligent video-surveillance systems to learn events of interest from a base of representative examples.

The work reported here uses the platform we defined in [3]. Such a system must, for example, include evaluation facilities, such as [4]. We report here an evaluation of the system for a wide range of sequences including CAVIAR [5] sequences.

The paper is organized as follows: section 2

describes the global system and its main characteristics; section 3 goes deeper in the understanding of the learning process of the vision system. Section 4 is devoted to performance evaluation of the system and section 5 concludes and indicates future work.

## 2. SYSTEM OVERVIEW

The global modular architecture system [3][6] is composed of heterogeneous computers and cameras connected together through a network (see Figure 1). A human computer interface and a storage device are also plugged onto this system. In order to test and benchmark our algorithm on particular recorded sequences, a general video stream player was added.



Figure 1. Global modular architecture.

### 2.1. Vision system description

The architecture of the vision part of the system is divided in three main levels of computation that achieve the interpretation (see Figure 2):

1. Image level (image filtering, background evaluation and segmentation)
2. Blob level (description, blobs filtering, matching, tracking description and filtering)
3. Event level (tracking analysis, sub-scenarios extraction and scenarios detection)

We will not go deeper in the explanation of the segmentation and tracking algorithm as it was described in [3]. It is a typical bottom-up approach.

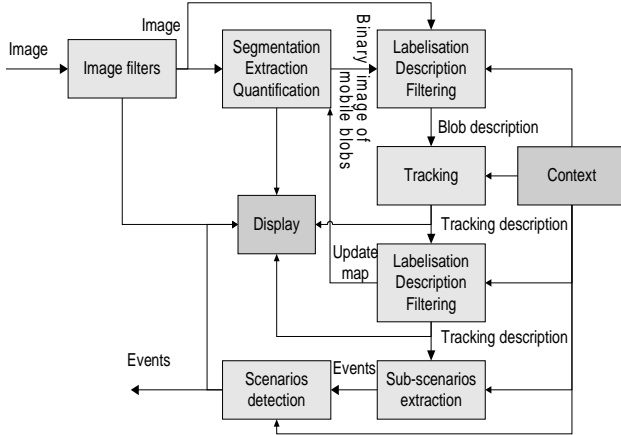


Figure 2. Design of the vision system components.

The output of the tracking blob level is a description of historical behaviours of blobs in the scene (see Figure 3). After sub-scenarios extraction, the description is in XML (see Figure 4).

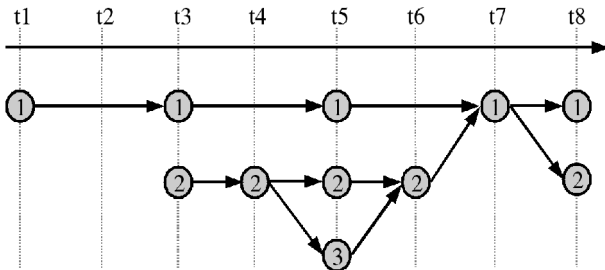


Figure 3. Historical behaviour of blobs in the scene.

```
<?xml version="1.0" encoding="UTF-8"?>
<event_history>
<event>
<time>t3</time>
<name>appearance</name>
</event>
<event>
<time>t4</time>
<name>Split</name>
</event>
<event>
<time>t6</time>
<name>Merge</name>
</event>
<event>
<time>t7</time>
<name>Merge-Split</name>
</event>
</event_history>
```

Figure 4. Description in sub-scenarios of scene chronological account.

### 3. LEARNING

We propose a new functional capability for our intelligent video-surveillance system to learn events of interest from a set of representative examples (see Figure 5). The basic idea is to use the Bayesian classification like in [7]. Let  $E$  be the evidence and  $H$  the hypothesis. Equation (1) is the rule to infer  $h$  knowing  $e$ .  $h$  is the expected scenario detection and  $e$  is the sub-scenario extraction.

$$P(H=h|E=e) = \frac{P(E=e|H=h)P(H=h)}{P(E=e)} \quad (1)$$

We employ the maximum *a posteriori* probability (MAP) selection rules [8]. When a feature  $e$  is received, the MAP system computes the *a posteriori* probabilities of  $P(h|e)$  and chooses  $h$  if  $P(h|e) > 0.5$ . From the set of representative examples, one can run the vision system description module (section 2) and compute the probability density function of  $P(E)$  with all the sub-scenario extractions  $e$ .

#### 3.1. Ground truth

The set of representative examples is composed of pairs of video sequences and expected results, known as ground truth. In our case, this ground truth is described in XML as shown in Figure 5. The structure of the ground truth is quite similar to the one of the blob description. With the ground truth, the probability density function of  $P(H)$  can be directly computed.

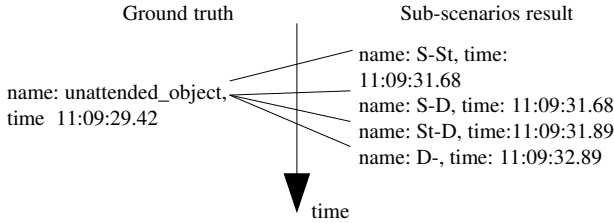
```
<?xml version="1.0" encoding="UTF-8"?>
<event_history>
<event>
<time>2003/08/08@17:35:12.520000</time>
<name>person_entering</name>
</event>
<event>
<time>2003/08/08@17:35:14.480000</time>
<name>unattended_object</name>
</event>
<event>
<time>2003/08/08@17:35:17.080000</time>
<name>person_leaving</name>
</event>
</event_history>
```

Figure 5. Typical ground truth of the representative learning base.

#### 3.2. Matching

It is possible to match the description in sub-scenarios  $E$  with the high level description of the ground truth. But several matches are possible. The bigger the time window,

the more the number of matches. We have decided to set the window to 5 seconds around the ground truth event. An example of matching is shown below. We match the ground truth `room2_00039.gt.events.xml` with the sub-scenarios results `room2_00039.events.xml`. There are 4 matches:



With all these matches one can easily construct the probability density function of  $p(E|H)$  event by event. Here the matching will increase the conditional probabilities  $p(S-St|unattended\_object)$ ,  $p(S-D|unattended\_object)$ ,  $p(St-D|unattended\_object)$ ,  $p(D-|unattended\_object)$ .

#### 4. EVALUATION

In public facilities such as airports, there is a risk of terrorist attack by abandoning a bag or an object containing a bomb. Detection in such a situation without delay might permit to avoid a disaster. We will evaluate the performance of our system with the case study of automatic detection of such unattended objects.

##### 4.1. Rationale

For objective performance evaluation we should define a dataset of sequences for learning, a dataset of sequences for testing and a metric to quantify the quality of the results. All sequences are downsized to 384x288 grey 5 fps and are acquired from fixed cameras. Table 1 defines the datasets. Sometimes, several people are on the scene. The ground truth has been made by hand and the structure is already defined in Figure 4. One can remark that the CAVIAR sequences are only used for testing the generalisation of the learning from other sequences.

The performance evaluation compares the result from the system and the ground truth. We use the metric described in [9] to characterize the successes and failures of the algorithm. Tables 2 and 3 remind the scoring process where  $N_{tp}$  is the number of observations confirmed by the ground truth.  $N_{fp}$  is the number of observations not matched in the ground truth.  $N_{fn}$  is the number of observations erroneously accepted as belonging to the ground truth.  $N_{tn}$  is the number of

observations rejected as belonging to the ground truth.

##### 4.2. Learning

The learning process gives  $p(h="unattended\_object"|e="S-ST")>0.5$ . For the other features, the probability is below 0.5. Thus, the benchmarking process will only assign as “unattended\_object” events when a sub-scenario “S-St” occurs. One can note that “S-St” represents the scenario of a blob splitting followed by a blob stopping.

“Multitel_test1”, 29 sequences, 22 unattended objects, $lr=15$ , $ts=14$	“Sas_cam2”, 43 sequences, 4 unattended objects, $lr=30$ , $ts=13$
	
	
“Multitel_outside”, 30 sequences, 2 unattended objects, $lr=15$ , $ts=15$	Caviar1[5], 4 sequences, 4 unattended objects, $lr=0$ , $ts=4$
	
	

Table 1. Description of datasets.  $lr$  is for the number of sequences used for learning process,  $ts$  is for the number of sequences for the testing process.

System Observation	Ground truth	
	Positive	Negative
Positive	$N_{tp}$ (true positives)	$N_{fp}$ (false positives)
Negative	$N_{fn}$ (false negatives)	$N_{tn}$ (true negatives)

Table 2. Boolean contingency table.

Name	Expression
Detection rate (sensitivity)	$N_{tp}/(N_{tp}+N_{fn})$
False positive rate	$N_{fp}/(N_{fp}+N_{tn})$

Table 3. Basic values for benchmarking.

#### 4.2. Results

Table 5 shows a detection rate of 72% and a false positive rate of less than 1%. This means that less than 1% of the negative event frame trigger a false alarm.

Number of event frames	Ground truth	
	Positive	Negative
14538	18	14520

Table 4. Description of ground truth of sequences.

Nb event frames	Ground truth		System observation			
	Positive	Negative	False positive $N_{fp}$	False positive rate	True positive $N_{tp}$	Detection rate
14538	18	14520	20	<1%	13	72%

Table 5. Results of the experimentation.

Despite the impression of good results, the quality of the system could be highly improved. This false positive rate (by frame) is still quite high. The results for the CAVIAR sequences show a detection rate of 50% that shows the results of generalisation for this event (unattended object).

After study, we discover that the problems are mainly coming from the segmentation of mobile objects and are then propagated along other processes (tracking and the sub-scenario detection).

### 5. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed an original approach for an automatic video surveillance platform that can learn events of interest. We have focused our attention to the vision system and demonstrated its capacity for automatic unattended object detection.

The results are quite interesting and show the relative ability to generalise learning to new sequences. We need to test other events different from « unattended object » that involve several people like « people fighting ».

Future work will aim at increasing the performance by overcoming different aspects of the system as segmentation, tracking but also by improving the learning.

#### ACKNOWLEDGEMENTS

This work has been granted by the Walloon Region under the FEDER project 171, the FIRST SPIN OFF program. We thank the reviewers for their helpful comments.

#### REFERENCES

- [1] Andrea Cavallaro, Damien Douxchamps, Touradj Ebrahimi and Benoit Macq: « Segmenting moving objects : the MODEST video object kernel », WIAMIS 2001, Tampere, Finland, May 16-17, 2001.
- [2] Cupillard F, Brémond F and Thonnat M, « Tracking groups of people for video surveillance », 2<sup>nd</sup> European Workshop on AVBS Systems.
- [3] Xavier Desurmont, Arnaud Bastide and Jean-Francois Delaigle, « A Seamless Modular approach for Real-Time Video Analysis for Surveillance », 5<sup>th</sup> WIAMIS 2004, April 21-23, 2004, Instituto Superior Tecnico, Lisboa, Portugal.
- [4] Jaynes C, Webb S, Steele R and Xiong Q, « An open development environment for evaluation of video surveillance systems », 3<sup>rd</sup> Int. Workshop on PETS.
- [5] the data as coming from the EC Funded CAVIAR project/IST 2001 37540, found at URL: <http://www.dai.ed.ac.uk/homes/rbf/CAVIAR/>.
- [6] B Georis, X Desurmont, D Demaret, S Redureau, JF Delaigle and B Macq, « IP distributed computer-aided video-surveillance system », Intelligent Distributed Surveillance Systems, IEE Visual Information Engineering Professional Network, 26 February 2003.
- [7] F. Lv, J. Kang, R. Nevatia, I. Cohen, G. Medioni, "Automatic tracking and labeling of human activities in a video sequence", Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, ECCV04. May 2004.
- [8] W.A. Varshney and P.K. Samarasekera, « A tight upper bound on the Bayesian probability of error », Dept. of Electr. Eng., Birzeit Univ., Pattern Analysis and Machine Intelligence, IEEE Transactions on, Feb 1994, On page(s): 220-224, Volume: 16, Issue: 2, ISSN: 0162-8828
- [9] Tim Ellis, « Performance Metrics and Methods for Tracking in Surveillance », Information Engineering Centre, School of Engineering, City University, London. Proceeding 3<sup>rd</sup> IEEE Int. Workshop on PETS, Copenhagen, June 1 2002.