# A VIDEO INTERPRETATION PLATFORM APPLIED TO BANK AGENCY MONITORING

B Georis M Mazière F Brémond M Thonnat

Inria Sophia-Antipolis, France

## ABSTRACT

In this article we present a real time platform for semantic video interpretation applied to bank agency monitoring. The proposed system is a multi-camera platform, which recognizes user-predefined scenarios, such as bank attack scenarios. These scenarios are modelled by domain experts using a back and forth process and based on a representation language. In order to address bank monitoring issues, the system has been improved based on two evaluation types. First, a repair stage guided by a careful technical evaluation has been performed at each level of the interpretation chain. As a consequence, the robustness obtained was sufficient enough to recognize all scenarios of interest. Second, an end-user evaluation has helped the experts to improve the scenario models to adapt them to real life situations. We report results of scenario recognition performances on real video sequences taken in a bank agency.

**Keywords**: scenario recognition, performance evaluation, semantic video interpretation

## 1. INTRODUCTION

In the past few years, many video interpretation systems have been developed in the computer vision community. Haritaoglu et al. (1) use shape analysis and tracking to locate people and their parts (head, feet,...) in image sequences. Oliver et al. (2) use Bayesian analysis to identify human interactions using trajectories obtained from a monocular image. Johnson and Hogg (3) have defined an efficient people tracker based on B-spline corresponding to people shape models. Nevertheless, few video interpretation systems have been successfully applied to real world applications due to a large variety of video interpretation issues. First, typical image processing problems come from shadows, illumination changes, over-segmentation or misdetection. Second, the tracking process remains a major issue. In video understanding, the loss of a tracked object prevents the analysis of its behaviour. In addition, most of these systems address vision issues and few of them provide a true semantic video interpretation. Hongeng et al. (5), Vu et al. (6) are part of the few examples able to perform complex reasoning (i.e., spatio-temporal reasoning) and to understand the interactions between people in real world applications. Finally, these systems usually perform well on a small video sequence set or in a well-constrained environment but results worsen in real conditions. Despite these facts, there is an increasing number of installed video surveillance systems being run 24 hours a day in varying conditions. Therefore, there is a strong need for highly reliable video interpretation systems with more and more reasoning capabilities. The recent creation of the PETS (Performance Evaluation of Tracking and Surveillance) workshops (4) shows the concern of the vision community to address this issue. It enforces the idea that evaluation techniques are needed to assess the reliability of algorithms.

In this paper we present a general-purpose activity monitoring platform. It can be used in various applications ranging from outdoor parking lot to metro station monitoring. We have chosen to focus on human behaviour recognition and to illustrate the platform through a bank agency monitoring application. To obtain a robust system adapted to a specific application, we propose to perform a repair stage guided by two evaluation processes. The first one is a technical evaluation. Its purpose is to correct vision algorithm errors by using in particular more contextual knowledge. The second one is an end-user evaluation. It allows us to refine scenario models to fit with real life situations.

The paper is organized as follows: section 2 presents the video interpretation platform, section 3 describes the scenario formalism and explains the scenario modelling, section 4 presents the evaluation and repair process together with results. Finally, section 5 concludes and gives future work.

## 2. VIDEO INTERPRETATION PLATFORM

Before going into details of the evaluation process, we present the generic and reusable activity monitoring platform (see Figure 1), which is able to recognize scenarios predefined by experts.

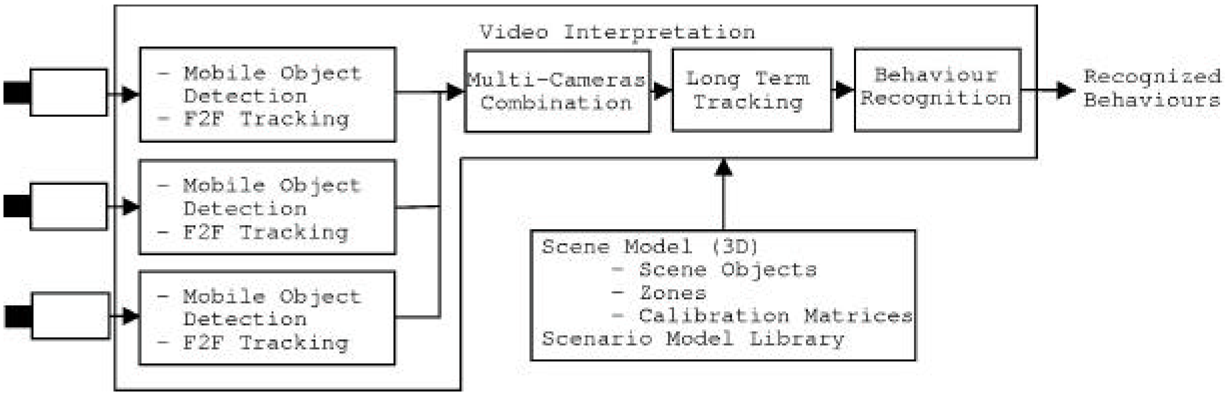

**Figure 1: Video interpretation process overview. It consists of five main functionalities: 1) Mobile Object Detection, 2) Frame to Frame Tracking (F2F Tracking), 3) Multi-Cameras Combination, 4) Long Term Tracking, 5) Behaviour Recognition. Contextual knowledge is provided for the whole interpretation chain.**

First, colour input images are digitized with a variable frame rate provided by one or several cameras. For each camera, a *mobile object detection* procedure detects moving regions by subtracting the current image from a reference image and classifies them into several predefined classes of mobile objects. Then, a *F2F (Frame to Frame) tracking* process links mobile objects over time and generates a graph of mobile objects for each camera. Nodes of this graph represent mobile objects while edges are temporal links over time. Second, a *multi-cameras combination* is performed to combine the graphs computed for each camera into a global one. Then, this global graph is processed by a *long term tracking* process for actors evolving in the scene. Its purpose is to robustly track individuals that correspond to a tracked person within a larger time window than the one of the F2F tracking. Finally, *behaviour recognition* is performed for each tracked actor. So, the output is a semantic description of the recognized behaviours. In addition, 3D scene models can be used as a priori contextual knowledge of the observed scene. For instance, a scene model can contain 3D positions and dimensions of static scene objects (e.g., a chair) and zones of interest (e.g., an entrance zone). Semantic attributes (e.g., fragile) can be associated to objects or zones to be used in the behaviour recognition process.

## 3. SCENARIO MODELLING FOR BANK MON ITORING

Scenarios to be recognized are the result of an interactive process with domain experts, which bring us important knowledge of human behaviours. Sub section 3.1 explains this process. The specific formalism used to describe such scenarios is then presented in sub section 3.2 and examples of them are given in sub section 3.3.

### 3.1. Domain Expert Knowledge and End-User Interaction

Many discussions with domain experts have been needed in order to define scenarios, corresponding to interesting human behaviours, which have to be recognized in bank agencies. A scenario can be thought of as two parts: the attack precursor part (i.e., the robber approach) and the attack part.

Today, classical bank agencies gradually evolve towards agencies with one or several counters without money, ATM (Automatic Teller Machine), safe room, and offices for commercial employees. The safe room is then the more significant zone inside the bank agency since all the available money is stored inside. As a consequence, all irregular behaviours or bank protocol infringement (involving either robbers or maintenance and cleaning employees) must be detected nearby the safe entrance. The protocol can be different for each bank. For instance, one of these rules is that only one person can enter the safe room. In this case, the system must raise an alert when more than one person is inside the safe room. For bank experts, this part of the scenario (people number inside the safe) must be recognized with a very high confidence.

Moreover, it is interesting to recognize a robber approach to the safe entrance or forbidden zones. Modelling all bank attack precursors is a difficult task due to their large number and variety. We list here some examples:

- counter attack: frequent, often stealthy, rapid and hardly observable even for human beings. The bank employee is threatened but it is generally difficult to see the difference with a classical customer request.
- safe attack: they are not frequent. Bank employees and customers are threatened. People are traumatised and things can take a bad turn.

- aggressive attack: bank employees and customers are threatened. The robber has lost his/her self control, money is not the main motivation and the robbery usually leads to a drama.

This scenario part is facultative for bank attack detection but important in order to anticipate potential actions and prevent any drama. Therefore, we have modelled a large set of scenarios to take into account the variety of bank robberies. Bank monitoring is a rich domain well adapted to assess the scenario modelling formalism. The next sub section describes the scenario formalism used during the modelling process.

### 3.2. Scenario Formalism

In order to describe scenarios, Vu et al. [6] have introduced a representation formalism which takes into account the expert knowledge. A scenario can be of different types and composed of states and events. A state is a spatio-temporal property defined at one time instant or on a time interval. An event is one or several change(s) of states at two successive time instants or on a time interval. Scenarios can be either primitive (single state change) or composite (combination of states and events). They are described by the following three parts:

- *physical objects* : all real world objects present in the scene observed by the camera.
- *components*: list of states and events involved in the scenario. They are facultative.
- *constraints* : all physical object, event or sub-event relations.

### 3.3. Scenario Examples

In a bank application, physical objects can be of two different types:

- *mobile objects* as people or group of people (robber, customer, kid, director, bank, maintenance, security or cleaning employee) and portable objects (suitcase, stroller or gun).
- *contextual objects* as predefined zones (entrance, back_counter, infront_counter, safe, safe_entrance) and equipment (counter, chair, desk, ATM, safe gate, poster, closet).

Currently, we have defined the following scenarios, with 1 to 3 persons (robber, bank employee, customer) moving on the 5 previous zone types and interacting with the safe gate equipment:

- *scenarios with 1 person* : the bank employee is behind or in front of the counter and goes to the safe. Then, the safe gate is opened.
- *scenarios with 2 persons* : the bank employee is behind the counter. The robber enters the bank agency, goes to the counter and stays in front of it. Both people go to the safe and the safe gate is opened.
- *scenarios with 3 persons* : the bank employee is behind the counter. A customer enters the bank agency, goes to the counter and stays in front of it. After, a robber joins the customer. The employee and the robber go to the safe and the safe gate is opened. The customer stays behind the counter or leaves the agency.

When constructing the scenario model library, we first select a set of primitive states and events. An example of each one is given in Figure 2 and 3 respectively. Composite events are then defined using this primitive set (Figure 4). In a second time, we can build more complex scenarios, which are a combination of primitive and/or composite events (Figures 5).

```
primitive_state inside_zone
  physical_objects: ( (p : Person), (z : Zone) )
  constraints: (p in z)
```

Figure 2: Primitive state model. The person *p* is inside zone *z*.

```
primitive_event changes_zone
  physical_objects: ( (p : Person), (z1 : Zone), (z2 : Zone) )
  components: ( (c1 : primitive_state inside_zone(p, z1))
                (c2 : primitive_state inside_zone(p, z2)) )
  constraints: (c1 before c2)
```

Figure 3: Primitive event model. Person p goes from zone z1 to zone z2.

```
composite_event Safe_attack_1person
  physical_objects:
( (p : Person), (z1: Back_Counter), (z2: Safe), (g:Gate) )
  components: (c1: primitive_event changes_zone (p, z1, z2))
  constraints: (g is opened)
```

Figure 4: Composite event with 1 person using the primitive event changes_zone. The person p goes from the zone back_counter to the safe zone while the safe gate is opened.

```
composite_event Safe_attack_2persons
  physical_objects:
    ( (employee: Person), (robber: Person), (z1: Entrance),
      (z2: Back_Counter), (z3: Infront_Counter), (z4: Safe) )
  components:
    ( (c1: primitive_state inside_zone(employee, z2))
      (c2: primitive_event changes_zone(robber, z1,z3))
      (c3: primitive_event changes_zone (employee, z2, z4))
      (c4: primitive_event changes_zone (robber, z3, z4)) )
  constraints:
    ( (c2 during c1)
      (c2 before c3)
      (c1 before c3)
      (c2 before c4)
      (c4 during c3) )
```

Figure 5: Composite event with 2 persons using the primitive inside_zone and changes_zone. It corresponds to the scenario with 2 persons where safe gate is closed.

## 4. EVALUATION CRITERIA AND RESULTS

Evaluation is realized at two levels. First, a technical evaluation enables to improve performances as it guides the repair stage performed at each interpretation module. Second, an end-user evaluation helps us to improve the scenario modelling step so that scenarios better correspond to real life situations. To perform the evaluation process, we need 2 types of data: video sequences and ground truth. More precisely:

- video sequences: various testing conditions must be considered to make a pertinent evaluation and to allow the subsequent repair stage. They must highlight different kinds of problems, which can arise at each level of the interpretation chain during a scenario recognition process. The technical evaluation has been realized on 2 video sequences from overlapping cameras, viewing the same scene in a cluttered bank agency during 400 frames. The scene contains 4 people crossing each other (2 bank employees, 1 robber and 1 customer) and 3 contextual objects (movable chair, counter and safe gate). 3 people are seen by the first camera and 4 people are seen by the second one. These videos are interesting since they highlight the chair displacement problem for the mobile object detection process, the frequent crossing of people for the tracking process and the complex composite scenario for the scenario recognition process. In addition, the system has been evaluated on hours of live and recorded video sequences without ground truth.
- ground truth: these data constitute reference data which are needed for the evaluation of each interpretation module. They must be defined as objectively as possible as shown in (Georis et al. (7)). Bounding boxes are drawn for each person even when he/she is dynamically or statically occluded in order to best fit the person. The only exception to this rule appears when the person is on the image border. The stored attributes are the 2D width and height, the 2D position and an identifier. In the next sub sections, overlap will refer to the percentage of a ground truth object covered by a detected mobile object.

Sub sections 4.1 to 4.3 describe the technical evaluation while sub section 4.4 describes the end-user one. The multi-cameras combination will be soon evaluated.

### 4.1. Mobile Object Detection Evaluation

Evaluation results are classified into three categories using mainly the overlap. A true positive corresponds to a high overlap, a false negative corresponds to a too weak overlap and a false positive is a detected bounding box not covered or not sufficiently covered by any ground truth object. We report 90% of true positives, 10% of false negatives, 3% of false positives for the first camera and 98% of true positives, 2% of false negatives, 15% of false positives for the second one.

This first evaluation has highlighted two main problems. The first problem was a false classification of persons when statically occluded. To lessen the false negative rate, we have thus improved the occlusion management due to 3D contextual objects by refining the projection of contextual and mobile objects onto the image to be able to correct the 3D parameter estimation of people. The second problem was a too high false positive rate. This was caused by a persistent change between the original background (empty scene) used for the background subtraction algorithm and the currently viewed scene. For instance, it was a new poster on the wall or a desk that was moved. So, we have added a noise tracking algorithm to discriminate between real moving regions and regions corresponding to a persistent change.

### 4.2. F2F Tracking Evaluation

For F2F tracking, classification into positive or negative tracks depends both on the overlap and on the presence of links between involved mobile objects. A true positive link is a link created by the tracking process combining two bounding boxes that both sufficiently cover a ground truth object at times t and t+1. All links made by the tracking process which are not true positives are classified as false positives. A false negative link is a link missed by the tracking process. We report 88% of true positives, 12% of false negatives for both cameras and 2% of false positives for the first one and 3% of false positives for the second one.

A typical problem emerged of the evaluation results for missing links. When a group of people was separating, the algorithm used to miss a link. To correct this frequent error, we have added a head detection procedure, which helps us to count how many people exist inside a group. This way, the tracker knows the number of links it has to create with bounding boxes corresponding to single person. In order to reduce the false positive rate, we have added a density criterion to avoid a link creation with a bounding box classified as a person while corresponding to a noise.

### 4.3. Long Term Tracking Evaluation

For the long term tracking evaluation, we have not used true positives, false negatives and false positives as they are not accurate enough. Instead, we have computed two metrics: the tracking time percentage per person and the number of identifiers per track. We report 91% of tracking time percentage, 3 identifiers per person on average for the first camera and 97% of tracking time percentage, 4 identifiers per person on average for the second camera.

People identifier switches are usually due to successive people crossings inside a small group. As only people staying close to each other switch their identifiers, the global tracking of the group is preserved and it does not prevent scenario recognition. Another limitation revealed by this evaluation was the incapacity of the algorithm to initiate an individual track where the individual was not located in an entrance zone. We have thus added a process to estimate the number of persons currently present in the scene in order to be able to initiate an individual track even if its corresponding person detection starts in the middle of the agency. For instance, this can occur when the person is dynamically occluded behind another person for a while.

### 4.4. End User Evaluation

The behaviour recognition evaluation has been realized in live condition inside a bank agency during one hour, together with end-users. A true positive corresponds to an alert raise when a real bank attack happens (simulated by actors), a false negative is the miss of an alert raise when a real bank attack happens and a false positive is an alert raise when no real bank attack happens. The bank_attack scenario with 3 persons was played 16 times. We obtained 93.75% of true positives, 6.25% of false negatives and 0% of false positives when 2 people enter the safe room. The scenario with 2 persons was played more than 10 times and we obtained 100% of true positives. The main reason why we obtained such true positives percentage is that scenarios were modelled with the interaction of domain experts through an incremental process.

## 5. CONCLUSION AND FUTURE WORK

In this paper, we have presented a semantic video interpretation platform for human behaviour recognition in bank agencies. This platform is already installed and tested in a bank. The system robustness is achieved through a repair stage based on a careful technical evaluation. This evaluation has been performed at each level of the interpretation chain. This repair consists in an efficient use of the contextual knowledge and not in major algorithm changes. This greatly improves performances. A second evaluation addressing end-user issues was carried out in live conditions with the presence of experts. It has first formally validated the scenario modelling step. Second, it has shown that the recognized scenarios were addressing in practice end user problems despite remaining tracking errors. The end users were able to play all scenarios in live and check directly the recognition results. Future work will investigate how we can automate the repair stage which is currently done by hand. Moreover, this application will be installed inside another bank agency in few months in order to check that the scenario model library we have defined is general enough.

**Acknowledgements:** thanks to B Macq and the Walloon Region for their funding for a part of this work. We also thank the bank agency of Crédit Agricole de la Brie for providing data and expertise.

## REFERENCES

1. Haritaoglu I, Harwood D and Davis L, 2000, “W4: real-time surveillance of people and their activities”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 809-830.

2. Oliver N, Rosario B and Pentland A, 2000, “A bayesian computer vision system for modelling human interactions”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 831-843.

3. Johnson N and Hogg D, 1996, “Learning the distribution of object trajectories for event recognition”, Image and Vision Computing, 14, 609-615.

4. IEEE Computer Society, Ed., 2002, “IEEE international series of workshops on Performance Evaluation of Tracking and Surveillance (PETS)”.

5. Hongeng S, Brémond F and Nevatia R, 2000, “Representation and optimal recognition of human activities”, IEEE Proceedings of Computer Vision and Pattern Recognition.

6. Vu T, Brémond F and Thonnat M, 2003, “Automatic video interpretation: a novel algorithm for temporal scenario recognition”, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI03).

7. Georis B, Brémond F, Thonnat M and Macq B, 2003, “Use of an Evaluation and Diagnosis Method to Improve Performances”, Proceedings of the Visualization, Imaging and Image Processing Conference (VIIP03).