PERFORMANCE EVALUATION OF EVENT DETECTION SOLUTIONS: the CREDS experience

F. Ziliani[1], S. Velastin[2], F. Porikli[3], L. Marcenaro[4], T. Kelliher[5], A. Cavallaro[6], P. Bruneaut[7]
[1]VisioWave, [2]Kingston University, [3]MERL, [4]Technoaware, [5]GE , [6]QMUL, [7]RATP
fziliani@ieee.org

Abstract

In video surveillance projects, automatic and real-time event detection solutions are required to guarantee an efficient and cost-effective use of the infrastructure. Many solutions have been proposed to automatically detect a variety of events of interest. However, not all solutions and technologies may satisfy all the requirements of the surveillance scenario. For this reason, performance evaluation of existing event detection solutions becomes an important step in the deployment of video surveillance projects.

In this paper, we propose a practical approach that aims at minimizing the ground truth generation problem and the expertise required to evaluate and compare the results by introducing specific requirements of specific event detection scenarios.. This approach is believed to be applicable for an initial evaluation of candidate solutions to a specific surveillance scenario before more exhaustive tests in an integrated environment.

The proposed method is under evaluation in the framework of the Challenge of Real-time Event Detection Solutions (CREDS).

1. Introduction

Object tracking, pattern recognition and in general image analysis and image understanding techniques offer today different approaches to automatic detection of events of interest in specific surveillance applications. Not all detection strategies behave equally well: a performance evaluation study is necessary to select the most appropriate solution to the specific problem.

The evaluation is often done directly by the users of the surveillance infrastructure. This is a long and expensive process that cannot be repeated for a large number of candidate solutions. In fact, system integration is needed before the evaluation is possible and users need to monitor the behavior of the automatic event detector for hours and days.

Although this approach guarantees that all the constraints of the surveillance scenarios are taken into account, it introduces a number of limitations:

- It is not suitable for comparing several event detection solutions on the same detection task and video sources.
- It requires time to guarantee the statistically completeness of the data: some events may be very rare.
- It requires an active feedback from the users who may not have the time to monitor performance continuously.
- It makes it difficult to learn from the observed limitations and update the solution accordingly.

An alternative and more systematic approach to the above strategy is to focus the performance evaluation effort on the modules constituting the event detection solution (e.g. the object tracking performances, the background update strategy, etc…). These approaches are successfully used by the academic community and enable the researchers to test, compare and improve specific image analysis modules on reference datasets.

Particularly popular are the initiatives that provide common datasets for evaluating object tracking or object segmentation techniques (e.g. PETS[1]). Since object tracking and segmentation provide the information that can be used to detect a broad class of events this approach can be a priori extended on several surveillance scenarios with different detection requirements. However, these strategies are driven by academic research and are often too generic to be successfully applied on real surveillance contexts. They require manual ground truth data generation, which is an extremely long and subjective process. Moreover, the evaluation results require a high level of expertise to be analyzed correctly. Finally, these approaches fail in providing a metric that enable easy

[1] www.visualsurveillance.org

comparison of results either against each other or against the ground truth.

In practice these modular approaches do not fully answer the industrial needs for performance evaluation. The main limitation is that the detection of events of interest is not necessarily correlated to the performances of the constituent modules used to achieve the detection. In some cases for instance, it is not necessary to have an extremely high performing object tracking to count people or vehicles in a scene. If we select the event detection only based on the quality of the tracking module we may chose the wrong solution. Moreover, these methods do not take into account the impact of the implementation or the way the modules are combined: all these choices are difficult to document and may make the difference between a good and a bad solution. Finally, these approaches cannot be used to compare methods based on completely different modules.

2. The CREDS approach

Based on the observations described in the previous section, an alternative approach has been investigated in the "*Challenge for Real-time Events Detection Solutions*" (CREDS) experience. The objective of CREDS is to propose and evaluate a practical intermediate strategy compared to those described in Section 1.The approach consists in the following procedure:

- The definition and description of a set of events of interest. These events are described precisely in order to reduce the efforts for manual ground truth generation.
- The collection of a realistic dataset. The dataset is composed of sequences with and without the events of interest.
- The manual definition of the ground truth for each sequence in the dataset.
- The definition of an absolute performance metric based on the ground truth data.
- The testing of each proposed solution using the dataset and the evaluation based on the performance metric.

The above procedure enables the direct comparison of event detection solutions based on any image analysis and understanding module. It takes into account the performances of the core modules by putting them in perspective of the usage scenario. The efforts to manually generate the ground truth data can be limited by simple and exact definition of the events to be detected.

The CREDS approach is intended to be used in a feasibility study to evaluate different solutions. Each solution can be tuned within this structure to achieve maximum performance. Once this phase is finished and the different strategies have been evaluated, the best performing ones may be integrated in the real architecture and undergo the final evaluation.

In order to evaluate the CREDS evaluation procedure, an open challenge for real-time event detection solutions has been proposed as part of the AVSS 2005 conference. The definition of the events, the datasets, the ground truth data and the metric has been proposed for a realistic event detection problem of the public transportation network of Paris (the RATP).

The information has been disclosed to the industrial and academic research communities and a call for solutions has been issued. An evaluation commission compares the solutions, the reported results and the computed metrics to select the best performing approach.

In the remainder of the paper, more details on the CREDS procedure are provided, whereas the final evaluation results will be presented during the AVSS conference by the evaluation commission.

3. Events definition

The CREDS dataset focuses on people safety in public transportation systems. In particular the sequences have been recorded by the RATP in a station of the Paris Subway network and show the same scene from different view points. Several scenarios have been recorded with different camera configurations. Each scenario consists of three video sequences corresponding to three video cameras in different positions (see Figure 1).

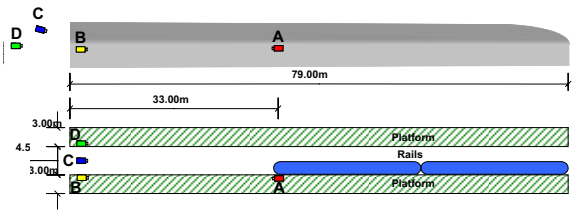

Figure 1: The topology of the camera configurations.

In this surveillance context we are interested in the detection of a number of events that if detected automatically can significantly improve safety in subway stations. These events are classified as warnings, alarms and critical alarms and are described in the following.

The warning events defined are:

- **Proximity warning**: A person stepping into the white line delimiting the platform or that extends any part of their body (legs, arms, etc…) over the rails even if they do not directly step on the white line.
- **Dropping objects on tracks:** A person dropping or launching objects from the platform to the tracks.
- **Launching objects across the platforms.** A person launching an object between the two platforms across the rails.

The alarms events defined are:

- **Person trapped by the door of a moving train:** A person that is trapped by the door of a moving train and is pulled along the platform.
- **Walking on rails.** A person walking parallel to the rails on the track.
- **Fall on the track.** A person crossing (toward the tracks) the white line delimiting the platform and falling or jumping on the rails area.

The critical alarm event is:

- **Crossing the rails.** A person crossing the rails from one platform to the other.

Each event is defined by an image zone and by an identification number (ID). The image zone is the spatial position identifier. By default this value is 0 (e.g. all the image), otherwise it is associated to one or more numbers, between 1 and 9, corresponding to specific zones in the image. All zones are identical and non-overlapping. The numbering and the zones are illustrated in Figure 2.

The identification number (ID) is composed of a start time and a duration. The start time is the time (in seconds) from the beginning of the input sequence, when the event is detected for the first time in an image zone. The duration is the time in seconds during which the event remains detected in the same image zone.

Figure 2: Image zones. The image is divided in 9 zones of the same size.

In a real environment, the automatic detection of the above defined events can trigger specific actions. Examples of actions are monitor redirections to pop up the specific video in the control room, storage, PTZ camera actions, audio signals, SMS, relays, etc. According to the type of events, different requirements are set concerning the false positive and false negative rates.

Table 1: Example of the ground truth data for the Scenario 3 and the Configuration 1. S represents the starting time in seconds from the beginning of the sequence, D is the duration in seconds, P1,P2,P3, are the positions in the image where the event has been detected, ID specifies the detected event.

S sec	D sec	P1	P2	P3	ID
5.64	5.20	5	5	2	1
10.84	0.40	5	5	2	6
11.56	7.28	5,6,9	5,6,8, 9	2	7
25.44	7.64	3	6	2	1
33.08	0.28	3,6	2,5	2	6
33.80	7.12	6,5	5	2	7
48.00	0.92	5,8	5,8	2	1
50.76	9.16	5,8	5,8	2	1
55.00	1.48	2	1	2	1
73.76	1.68	2	1	6	1
75.44	0.12	2	1	6	6
75.84	6.64	2,3	1,2,5	6,5, 4	7
88.16	2.84	3	2	-	1

4. Ground truth

Using the event definitions given in Section 3, it is relatively easy to manually generate the ground truth data. Each sequence in the dataset is manually annotated and a ground truth description has been provided for each surveillance scenario. The ground truth data has been made available on the conference FTP[2] together with the corresponding dataset.

An example, of ground truth data is shown in Table 1. Similar tables have been provided for all the 12

[2] www-dsp.elet.polimi.it/avss2005

scenarios organized in 4 camera configurations. Note that the efforts required to draw such ground truth are quite limited compared to those necessary in more general frameworks such as object tracking. This is an important advantage in industrial applications.

5. Metric

The ground truth data is used to measure the deviation of the automatic event detection solution from the ideal behavior. To this end, a metric is defined to compute a score. Large positive scores identify good performing solutions, whereas large negative scores identify bad performing solutions. The same metric can be used to compare different results.

The CREDS metric first defines correct detections, false positive and false negative detections. Then, it associates to each of these detections a weight. The final score for a given scenario is the sum of all the corrected, false positive and false negative detection scores.

A correct detection is defined as the first occurrence of an Automatically Detected Event (ADE) that overlaps in time a Ground Truth Event (GTE) with the same ID. We identify three different kinds of correct detections, namely perfect, anticipated, and delayed.

- **Perfect**: the GTE and the ADE are the same.
- **Anticipated**: the ADE starts before the GTE.
- **Delayed**: the ADE starts after the GTE.

Note that each GTE may be associated to a single correct detection. If multiple ADE overlaps in time the same GTE, only the first overlapping ADE will be considered as a correct detection. The others will be classified as false positive detections. A false positive detection[3] (false alarms) correspond to automatically detected events (ADE) that do not overlap in time any of the not yet associated ground truth events sharing the same ID. False negative detections[4] (miss detections) correspond to ground truth events (GTE) that do not overlap in time any automatically detected event sharing the same ID.

The score for correct detections, $S_{CD,}$ is a function of the delay/anticipation, t, and of the ratio between the GTE and ADE durations. The score is computed with the following function:

$$S_{CD}(t, ADE_{duration}, GTE_{duration}) = \begin{cases} 0 & t \in]-\infty, B[, t \in]D, \infty[; \\ \frac{S}{A-B}(t-B) & t \in [B, A]; \\ S & t \in]A, 0[; \\ \frac{S}{D}(D-t) & t \in [0, D]; \end{cases}$$

This is represented graphically in Figure 3.

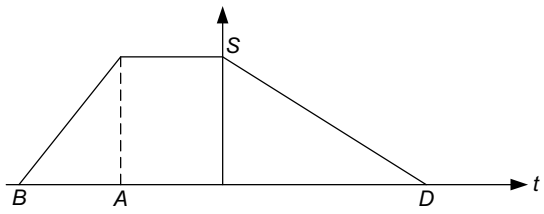

Figure 3: The function used to measure the score in case of correct detections.

The X axis represents the anticipation/delay time in milliseconds: positive values are associated to a delay and negative values are associated to anticipation. The definition of S, B, A, and D is provided below.

S represents the maximum score associated to a correct detection: it is computed as:

$$S = \begin{cases} 50 * \left[2 - \left(1 - \frac{ADE_{duration}}{GTE_{duration}} \right)^2 \right] & \frac{ADE_{duration}}{GTE_{duration}} \in [0,2] \\ 50 & \frac{ADE_{duration}}{GTE_{duration}} \in]2, \infty[\end{cases}$$

S belongs to an interval between [50, 100] and its graph is represented in Figure 4.

D represents the maximum delay in ms tolerated for a specific event: a delay above D is associated to a null score. A represents the accepted anticipation in ms: a detected event with an anticipation smaller than A is associated to a score equal to S. B represents the maximum tolerated anticipation in ms: anticipation above B is associated to a null score.

The values for D, A, and B are different according to the type of event (Warning, Alarm or Critical Alarm).

There are two different constant scores associated to the two different types of false positive detections. The first S_{FD1} is associated to the one which does not overlap any GTE, and the second S_{FD2} is associated to the one which overlaps a GTE already associated to an ADE. The values of these two constants are different according to the type of event (Warning, Alarm or Critical Alarm).

[3] A **false positive** exists when a test reports, incorrectly, that it has found an event where none exists in reality.

[4] A **false negative** exists when a test reports, incorrectly, that an event was not detected when, in fact, was present.

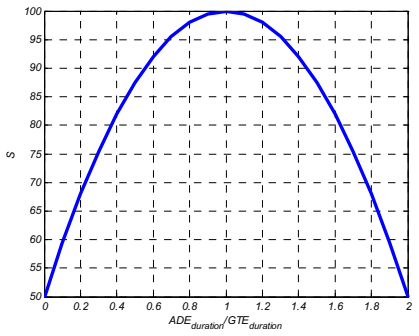

Figure 4: Score as a function of events duration.

The score associated to false negative (miss) detections S_{MD} is a constant. Its value depends on the type of event (Warning, Alarm or Critical Alarm).

Based on the above definitions we propose three different measures, the first for warning events, the second for alarm events and the third for critical alarm events. The measures are different because the detection requirements associated to warnings, alarms and critical alarms differ. When warning events are detected by automatic image analysis, specific actions need to be triggered such as starting a recording, broadcasting an audio message or redirecting input video on a specific monitor for human verification. In the case of warning events, it is acceptable to tolerate detection delays and or anticipations of the order of a few seconds. False negative detection and false positive detections are not desirable, but clearly not critical. In addition to the detection of the event, the detection of its position in the image can be used to associate the same type of event to different actions or to highlight where in the image the event has been detected.

When an event classified as an alarm is detected, similar actions as for the warning events need to be triggered but constraints on delay, anticipations and false negative detections are higher. These constraints are even stronger in case of critical alarm events where any false negative detection may cause human casualties and any false positive detection heavy expenses.

Table 2: The parameters chosen in the evaluation metric to take care of the requirements of each class of events.

	A ms	B ms	D ms	S_{MD}	S_{FD1}	S_{FD2}
Warnings	-1000	-2000	2000	-100	5	-50
Alarms	-1000	-2000	1000	-250	-5	-80
Critical Alarms	-1000	-2000	200	-1000	-5	-200

In case the position of a detected event is not correct, the score is reduced of 10%. In case the position is set to 0 (the whole image), the score is reduced of 5%.

6. Proposals submitted to CREDS

Despite the novelty of this evaluation procedure and the limited time framework available to submit a proposal, the CREDS initiative has raised interest in both the academic and industrial communities. Four final candidates have successfully submitted their solutions to the specific event detection tasks described in Section 2 and have accepted to follow the performance evaluation procedure described in Section 4. The four proposals are described in detail in [1], [2], [3] and [4].

Each proposal adopted a different strategy to detect the events of interest as described by the application requirements defined by the RATP. Some of them used object tracking modules, others choose local motion estimation approaches instead, and most of them used a background extraction technique.

In [1], a non linear background update strategy has been adopted followed by a morphological analysis and a region growing segmentation. The generated blobs are characterized through feature extraction and a tracking mechanism is used. The model associated to each target is progressively updated and classification enables to identify if the target is a person or an object. The analysis of the detection and tracking results and the configuration of the scenarios enable the automatic detections of warnings, alarms and critical alarm events.

In [2], a block based motion estimation approach is used in a statistical framework to robustly generate background estimation. Foreground segmentation benefits from the robustness of the background and enables the approach to detect several events of interest. The configuration of the ground plane introduces perspective information while the definition of region of interest enables the triggering of specific event detection.

In [3], a low level analysis based on feature extraction and tracking is followed by a high level analysis which introduces the scenarios configurations to detect the events of interest. The low level analysis combines several modules such as motion detection, background differencing, and histogram analysis. The high level analysis is based on principle of artificial intelligence, and is used both for target tracking and classification.

Finally in [4], a 3D calibration of multiple cameras has been used to generate perspective information.

Background analysis followed by region growing segmentation of moving objects and model-based tracking detect the targets in the scene. These are compared with the requirements of the scenario and of the events to be detected to provide the final detection results.

A number of observations can be derived from these solutions. It is interesting to note that the solutions have not used all the available video sequences. Some approaches preferred the sequences generated by the camera at the top of the tunnel; others used more conventional cameras positions. Only one approach used multiple cameras in the same scenario to perform the detection. This may be explained by either the limited time available to participate to the challenge or the wide base-line approach that has not yet proved its benefits in improving the image analysis results.

Moreover, it is also worth to note that most approaches offer a general detection framework that can be configured and tuned to comply with the requirements of the RATP problems, but that can also be applied to other conditions and to other event detection tasks. The modularity of the solutions is at the basis of all the proposed approaches.

Finally several authors underlined the need for robust detections: one of the principal requirements of any automatic detection system is to be able to run on a 24/7 basis. If the conditions for a reliable detection are not met the system is required to notify its limits and prevent false positive and negative alarm rates to increase. The current evaluation procedure cannot be used to assess such robustness due to the limited amount of data available.

7. Conclusions

We described the Challenge of Real-time Event Detection Solutions (CREDS) organized during the AVSS 2005 conference. Specific event detection scenarios and requirements have been proposed by the public transportation company of Paris (RATP) and VisioWave. These scenarios focus on the detection of warnings, alarms and critical alarms in subway stations.

The challenge proposed a specific performance evaluation procedure that can be used in practical projects to identify among the candidate solutions those which satisfy the event detection requirements most. The proposed performance evaluation is based on a simplified definition of ground truth data and on an objective performance measure.

Four original strategies have been submitted to the challenge and are published in the AVSS 2005 proceedings. The results of these strategies are currently being evaluated according to the proposed methodology by an evaluation commission of experts. After the completion of this task, a report on the advantages and limitations of the CREDS evaluation procedure together with a comparison of the different solutions will be presented at the AVSS conference.

References

[1] M. Spirito, C. S. Regazzoni, L. Marcenaro, “Automatic detection of dangerous events for underground surveillance”, in Proceedings of the IEEE AVSS, 15-16 September, Como, Italy.

[2] J. Black, S. Velastin, B. Boghossian, “A Real Time Surveillance System for Metropolitan Railways”, in Proceedings of the IEEE AVSS, 15-16 September, Como, Italy.

[3] K. Schwerdt, D. Maman, P. Bernas, E. Paul, “Target Segmentation and Event Detection at Video-rate: the EAGLE Project”, in Proceedings of the IEEE AVSS, 15-16 September, Como, Italy.

[4] C. Seyve, “Metro Railway Security Algorithms with Real World Experience Adapted to the RATP Dataset”, in Proceedings of the IEEE AVSS, 15-16 September, Como, Italy.