TEC2011-25995 EventVideo (2012-2014)

Strategies for Object Segmentation, Detection and Tracking in Complex Environments for Event Detection in Video Surveillance and Monitoring

D3.2

EVENTS DETECTION IN DENSE ENVIRONMENTS

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by

AUTHOR LIST

Fulgencio Navarro Fajardo	fulgencio.navarro@uam.es
Diego Ortego Hernández	diego.ortego@uam.es
Juan Carlos San Miguel Avedillo	Juancarlos.sanmiguel@uam.es

CHANGE LOG

Version	Data	Editor	Description
0.0	11-11-2014	Diego Ortego Fulgencio Navarro	Initial version.
0.1	15-11-2014	Diego Ortego	2.2.1.1, 2.2.1.2 and 2.2.4
0.2	21-11-2014	Fulgencio Navarro	2.2.2
0.3	08-12-2014	Juan Carlos San Miguel	2.2.1.3 and 2.2.3
0.4	18-12-2014	Fulgencio Navarro Diego Ortego	Review
1.0	20-12-2014	José M. Martinez	First version

CONTENTS

1. Introduction

In recent years, there has been an increasing necessity of security in public and private facilities. This has led to the great importance that automatic video-monitoring has nowadays. Among the automatic video-monitoring applications, those that aims to detect dangerous events of different nature such as anomalous behaviors, abandoned and stolen objects or fall detection, have become an important research topic due to their social utility.

An important research line in this field is the anomaly detection. This task is becoming relevant as a preliminary stage in more complex event detection algorithms, and also as a simple alert application in video-surveillance. It is a dependent task application based on defining normalcy models.

Detecting stolen and abandoned objects is a relevant task in the video-surveillance field, and there are numerous and different approaches in the state-of-the-art. However, proposed solutions are not of wide application. The state-of-the-art is working on including new features to overcome the problems related to heavy occlusions or illumination changes and to provide long-term algorithms.

These two tasks, and many others, can be englobed in action recognition algorithms. However, even if the video-surveillance future seems to go in this way, the current approaches still presenting poor results, due to the high level of complexity required in this algorithms. An example of action recognition that has become of interest in the independent-living scenario is the fall detection.

In this document we summarize the works developed covering the aforementioned research areas and we further include two works which, besides coping with event detection, they propose two frameworks for smart and long-term video-monitoring, based on feedback strategies and context modelling.

1.1. Document structure

This document is composed of the following chapters:

Chapter 1: Introduction to this document.

Chapter 2: Contributions

Chapter 3: Conclusions and future work.

2. Contributions

This chapter compiles the contributions developed within the scope of this project.

2.1. Anomaly Detection

In this Master thesis [1], a comprehensive study of the existing anomaly detection framework has been carried out. The detection of anomalies in video surveillance sequences has gathered considerable interest as a research topic in the recent years. Traditionally, researchers have taken a pattern recognition approach to detect a set of previously defined events. However, these approaches are generally limited to constrained scenarios and cannot be easily generalized for arbitrary behavior. More recently, there has been a paradigm shift towards statistically modelling normal behavior in a scene, focusing on detecting behavior what stands out from the surrounding context. The contributions of this work can be summarized as follows:

- the identification of current challenges in anomaly detection. An extensive study of the state-of-the-art has been carried out to identify key challenges in anomaly detection. In order to compare existing approaches, three key areas have been identified: definition of anomaly, extracted features, and evaluation method.
- the implementation and evaluation of an existing framework for anomaly detection. An existing approach from the literature has been selected and implemented.
- the proposal of improvements to the base algorithm for challenging scenarios.

Anomalies can be broadly defined as an observation that stands out from the surrounding context (see Figure 1).

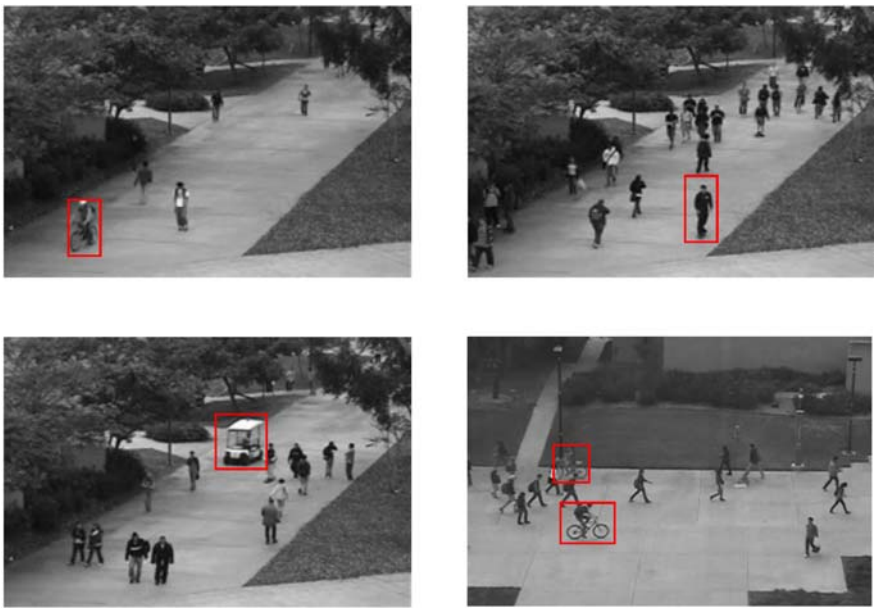


Figure 1. Sample anomalous events from UCSD Anomaly Detection Dataset.

While intuitive, this definition has led to subjective interpretations of anomalies, which have resulted in very diverse approaches aimed at solving the same problem. Depending on the nature of the features extracted to model normal behavior and anomalies, different anomalies can be considered. This has made it difficult to compare existing techniques, as they often rely on different definitions for anomalies, and can result in different authors identifying different anomalies on the same datasets. Additionally, the infrequent nature of anomalous events makes

then infrequent in current video datasets, and most authors evaluate their approaches on a very limited number of video sequences.

We can distinguish between pixel-based approaches, where features are extracted at pixel level, and object-based approaches, where features are associated with an object or blob. Among pixel-based ones, we can find a wide variety of extracted features such as pixel change frequency and pixel change retainment to capture spatio-temporal behaviors, filling ratio of foreground pixels, histogram of pixel change frequency, gradient magnitude, accumulation of pixel differences and optical flow. Among object-based ones, they can either be derived from appearance features (blob size and texture) or motion features derived from tracking (blob speed, direction and orientation).

However, while anomalous events are easy to define intuitively, there are a number of factors that pose challenges to anomaly detection techniques; the detection of anomalies is heavily dependent on how normality is modelled and which features are extracted (context and scale); a non-stationary context may alter normality at different times in a given scenario; anomalous events are generally infrequent, sparse, and unpredictable leading to a limited number of examples in training sequences, thus turning validation of techniques into a challenge task.

Attending to the evaluation methods, in order to perform validation, an anomaly detection method must be tested on a data set of test sequences that contain instances of previously annotated anomalies. However, this process poses different challenges. Firstly, the concept of anomaly varies on each approach depending on the features extracted and whether or not contextual information is employed, which may even result in different anomalies being defined on the same datasets. Therefore, establishing a ground truth is heavily reliant on subjective perception. Secondly, the availability of anomalies in video datasets is scarce due to their infrequent nature, making it difficult to provide statistically significant performance metrics. Due to these challenges, some authors have had to provide subjective performance assessments. In other works, authors manually annotate sequences from video datasets in order to provide performance metrics (precision/recall). For approaches that consider the detection of anomalous trajectory paths, the work is simplified by annotating the ground truth of extracted paths in test sequences. However, there is no unified criterion on which paths have to be considered anomalous, and in some cases authors do not provide criteria at all. Sometimes, ground truth is provided by different subjects, while approaches that perform clustering of trajectories simply consider clear outliers are anomalies. For approaches based on pixel-level abstractions, ground truth becomes more difficult to elaborate as it should label anomalous pixels in individual frames. Moreover, some works define different level in the ground-truth, namely, frame level and pixel level. For pixel level evaluation, localized detections are compared to ground truth masks. A correct detection is considered if at least 40% of anomalous pixels are labeled correctly. Other authors have followed the same evaluation framework, and have been able to provide performance comparisons of different approaches on the same datasets.

In this work we have implemented an existing approach, from now on base algorithm, from the literature [2]. This technique describes a framework that is capable of modelling normal activity in the scene. For this purpose, a "background behavior image" that captures background activity in the scene is constructed from training data. Activity in the scene is modelled at the pixel level by extracting features from regions in the image that are determined to be in motion. In order to detect anomalies, an image that captures current activity with associated features is constructed. Anomalous events are detected by comparing this image to the background behavior image. In this work, a fixed camera is assumed. Additionally, the authors impose the requirement of temporal stationarity of normal activity in the scene. Normal activity is defined as motion that is considered normal in the scene, which includes certain phenomena such as fluttering leaves in the background, moving water surfaces, or regular motion introduced by camera vibration. An overview of the system is shown in Figure 2.

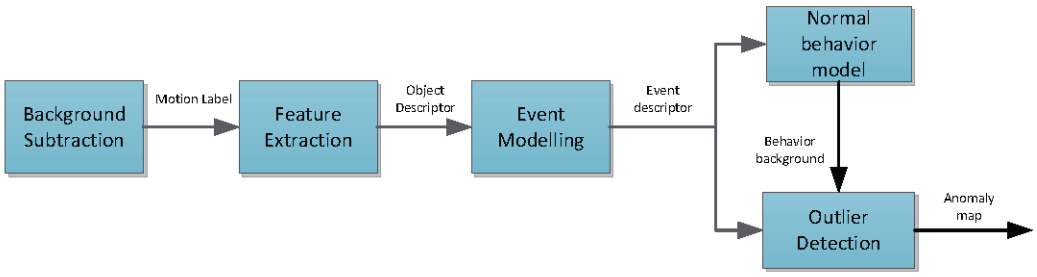


Figure 2. Anomaly detection framework.

At an initial stage, frames are captured from a static camera. For each frame, activity is characterized by labelling each pixel as either "moving" or "static". This motion label image can be computed employing existing background subtraction techniques. In order to characterize the motion occurring at each pixel location, a pixel-level behavior signature image is then computed. This behavior signature consists on a feature descriptor that can include features such as the size, shape, speed and direction of objects passing through individual pixel locations. In the event modelling stage, events modelled using a 2-state Markov chain. Events are defined as the behavior signature (represented by the feature descriptor) left by moving objects over a time window. In this period of time, the pixel goes through transitions between the two aforementioned states, moving or static. In the training phase, the event signatures of normal activity are employed to construct a behavior background image that depicts normal activity in the scene. For anomaly detection, extracted events are compared against the behavior background to provide an anomaly map, depicting the location of anomalous motion.

Furthermore, we have introduced improvements to the base algorithm for challenging scenarios by including a modified size descriptor that is invariant to spatial scale, thus improving the system for the detection of anomalous behavior of small objects (see Figure 3).

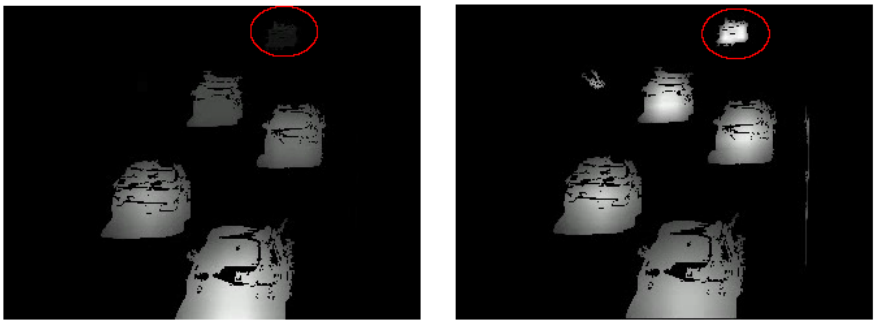


Figure 3. Comparison between size descriptors of base algorithm (left) and proposed algorithm (right).

Additionally, motion features have been included in order to detect anomalies caused by object motion in irregular directions.

In future work, the elaboration of a comprehensive dataset for evaluation and validation of anomaly detection will be done, using longer video sequences to properly model normal behavior. Also speed-related features will be explored to make the system robust against anomalies due to unusual speeds. Moreover, the elaboration of strategies to provide robustness against changing context will be performed, as most approaches work under the assumption of a "stationary normality" which is often unrealistic, as motion patterns of subjects and objects is often conditioned on a context that can change in time.

2.2. Activity Detection

2.2.1. Stolen and Abandoned Object Detection

In the video surveillance domain, the automatic detection of abandoned and stolen objects in real-time has recently become a topic of great interest especially in crowded environments. In general, this detection is achieved by developing a system with the following analysis stages: foreground segmentation, stationary region detection, blob classification and abandoned/stolen discrimination. The last stage of this pipeline determines the system's ability to discriminate stationary foreground objects (SFO) between abandoned and stolen. The contributions developed in this research area are related with the SFO detection (2 papers) and classification of those SFO into stolen or abandoned (1 paper).

2.2.1.1. Stationary Foreground Detection using History Images

Detecting SFO is an active area of research in many video-surveillance areas such as the detection of abandoned objects and illegally parked vehicles. Stationarity is defined as an object, person or group of people remaining stopped after previous movement. This task remains unsolved for complex sequences such as crowded scenarios as it faces many challenges related with illumination changes, low resolution images, object occlusions, high density of moving objects (increasing the number of cast shadows) and initialization of the detection algorithms. In this paper [3], we propose an approach for SFO detection in video based on the spatio-temporal variation of foreground and motion data (see Figure 4). Two parallel analysis are performed to segment both Foreground and Motion data and subsequently compute their History Images. Then, a Combination together with an occlusion handling method is applied to obtain the result, i.e. the Static Foreground Mask (SFG) with the SFO.

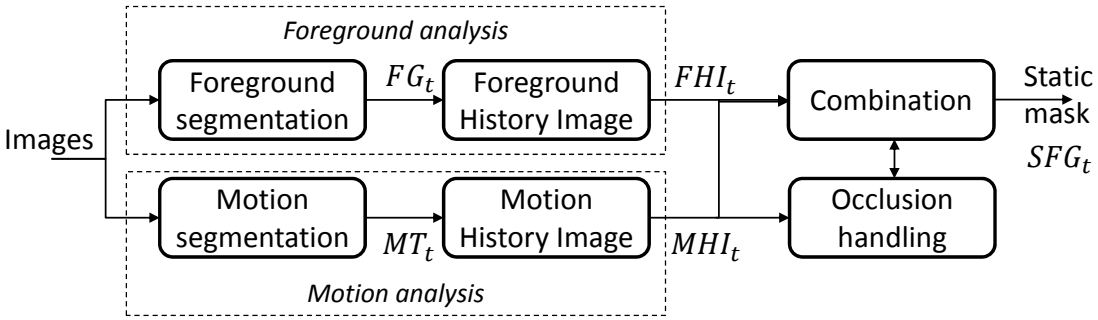


Figure 4. Overview of the proposed approach.

Deeping in the proposed approach, the first step is to segment Foreground and Motion data. Foreground data are obtained by Background Subtraction to detect regions of interest, while Motion data allow to filter out the moving regions and it is estimated using a novel technique which is based on median filters over sliding windows. We regarded the use of motion information in recent works for filtering false positives caused by high densities of moving objects. However, the extended motion information used, based on thresholding inter-frame differences, is not able to distinguish stationary objects continuously occluded, i.e. to detect no-motion behind motion. To tackle the aforementioned issue, we propose extending the motion analysis over temporal windows of length T (see Figure 5) as although multiple occlusions affect stationary regions in crowds, they usually last for few frames and the most predominant region in short-time intervals corresponds to the stationary one. For extracting motion using temporal windows, we apply a median filter before and after the frame under analysis, thus obtaining two median images that are subtracted and then thresholded to obtain the motion data.

The second step in both the Foreground and the Motion Analysis is to compute spatio-temporal information of the segmented data. To that end, we use History Images, which consist in the accumulation of the segmented data over time, thus computing the Foreground History Image (FHI) and Motion History Image (MHI). Then, both History Images are combined using the average to model the stationarity over time. Finally, the static detections are obtained using a two threshold scheme that considers motion activity.

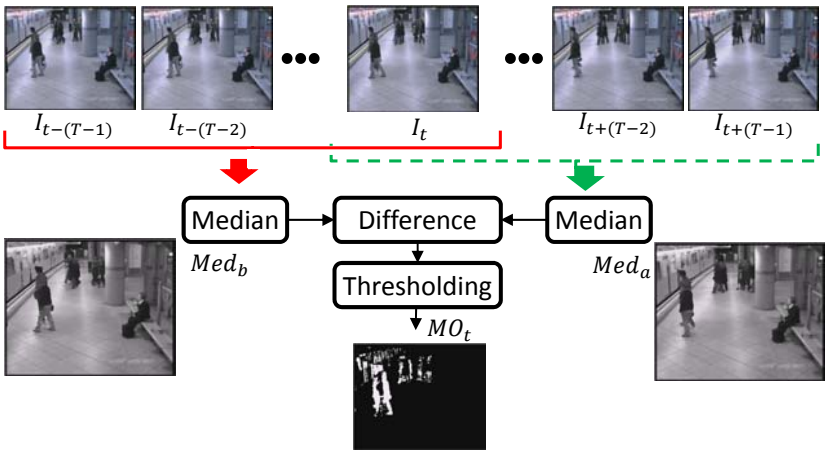


Figure 5. Motion segmentation scheme.

We compare the proposed approach with the most popular approaches based on foreground accumulation [4] (Acc), subsampling [5] (Sub) and foreground-motion sampling [6] (Bay). Unlike previous work, the proposed approach is able to maintain the SFO detection rate while removing false detections caused by high motion. Additionally, Figure 6 presents a comparison of the evaluated approaches where it can be seen how the proposed approach removes false detections due to motion activity while keeping the detection of the suitcase. For the evaluation, a ground-truth of SFO was annotated.

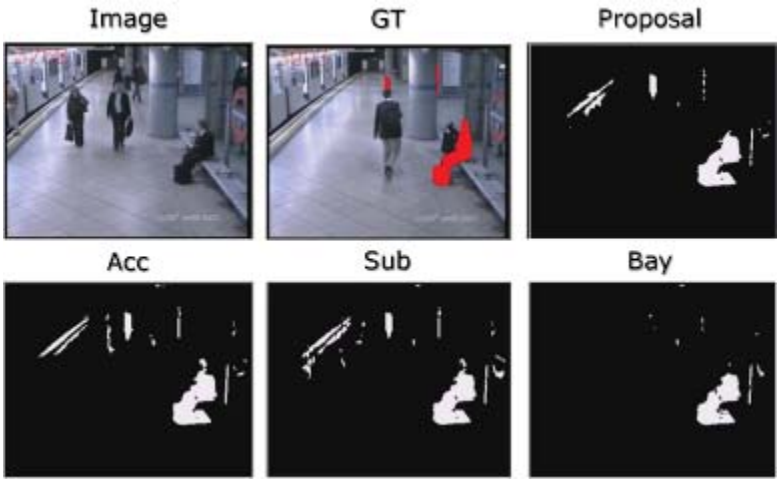


Figure 6. Comparison of static masks.

2.2.1.2. Multi-feature Stationary Foreground Detection for crowded environments

In this paper [7], we extend our work for SFO detection presented in [3] by formalizing its two-feature combination into a generic framework to combine multiple features (see Figure 7). The proposed scheme, unlike state-of-the-art approaches, jointly tackles several limitations, namely, continuous occlusions, high dense environments, shadows and illumination changes, thus leading to an approach suitable to operate in crowded environments. In this framework, for each feature, two common stages take place: Feature Map (FM) extraction and History Images (HI) computation. Then the Combination & Thresholding stage combines the feature results to obtain the Stationary Foreground Detection mask.

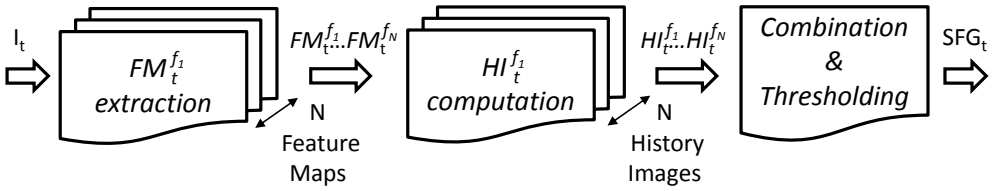


Figure 7. Overview of the proposed approach.

Deeping in the proposed approach, we employ three features: Foreground and Motion introduced in [7] and, as novelty, a Structural feature based on the Structural Similarity (SSIM) Index [8]. The main features used for stationary detection (foreground and motion) do not handle illumination changes, limiting their efficiency. Therefore, we propose to address such problem via a structural based feature. We use SSIM, originally developed for image quality assessment (i.e., between modified and distortion-free images), returning value 1 for highest quality. SSIM compares two images using three components; luminance, contrast and structure and computes every component over each pixel neighborhood, thus providing a SSIM map at pixel level. We obtain this SSIM map for SFO detection by comparing the current frame and a background model. Such comparison determines which pixels belong to object (or background) due to their different (or equal) structure to the background model. Figure 8 shows that SSIM identifies shadows and illumination changes areas as background, having high scores when comparing current frame and background:

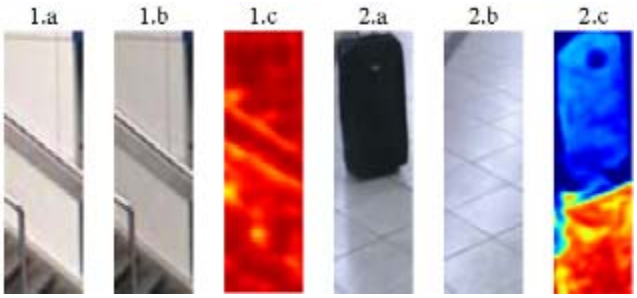


Figure 8. SSIM map (c) between frame (a) and background (b) patches, where the examples are: (1) an illumination change and (2) an object with its shadow. Dark blue (red) refers to min (max) SSIM scores. In example 1, SSIM (1.c) has high similarity scores despite the different illumination of frame (1.a) and background (1.b). In example 2, SSIM (2.c) has high (low) values in the shadowed (suitcase) area when comparing frame (2.a) and background (2.b).

The proposed structural information feature, Region SSIM map (RSIMM), is the mean of the SSIM map over a square window of size Q×Q centered at the pixel under analysis (minimum

score is bounded to zero). The mean operation is applied to handle the performance decrease of SSIM when after an illumination change, the color is locally saturated. The higher Q, the higher the robustness against saturation but the lower the precision.
Having the three FM computed, three HI are obtained and combined to model spatio-temporal stationarity. Such combination is thresholded taking into account motion activity and using an occlusion handling method.

We compare the proposed approach with the popular state-of-the-art approaches presented in [3] and we extend such comparison including the approach [9] (Dual) and our previous approach [3] (Med). Additionally we extend the amount of evaluation sequences. We achieve a higher performance than selected state-of-the-art, especially in crowds. Such enhancement is due to the high reduction of false detections in cases of shadows and illumination changes while keeping the correct detections. Additionally, Figure 9 presents a comparison of the evaluated approaches where it can be seen how the proposed approach removes false detections due to motion activity and shadows and illumination changes while keeping the detection of the suitcase.

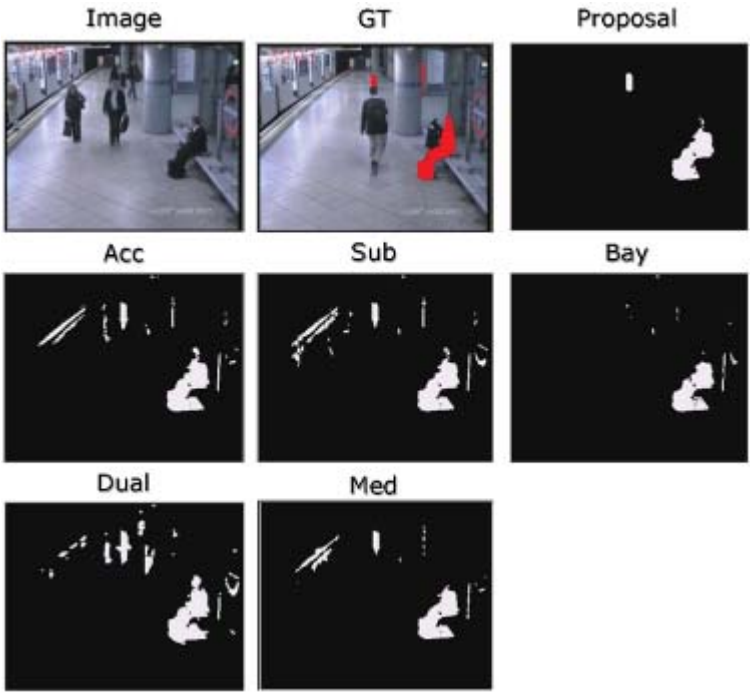


Figure 9. Comparison of static masks.

2.2.1.3. Pixel-based color contrast for stolen and abandoned object detection

In this work, we focus on the automatic detection of abandoned and stolen objects in real-time, i.e. in the last stage of the pipeline introduced in 2.2.1 which purpose is to determine the system ability to discriminate stationary foreground objects between abandoned and stolen. For its implementation, the common approach is to study the similarities between features extracted from the current and background frames of the video sequence.

We investigated a new approach for discriminating stationary objects into abandoned and stolen by using the color contrast along the object contour at pixel level. It assumes that object contour coincides with the color boundaries of the frame. Opposed to current approaches, it does not require specific background properties being suitable for complex backgrounds and

non-accurate foreground segmentation masks allowing real-time operation. The block diagram of the proposed discrimination scheme is depicted in Figure 10.
Computational cost results show that our approach highly reduces the time execution. A reduction factor higher than 92% was achieved.

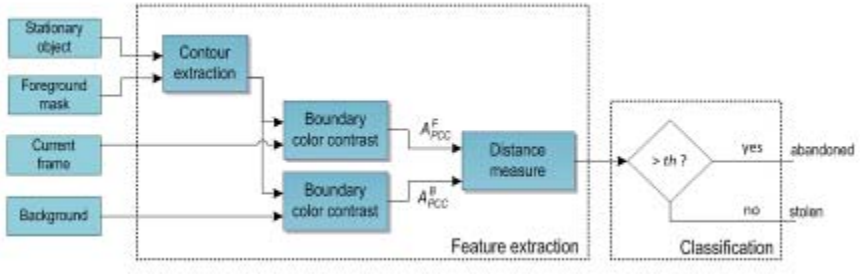
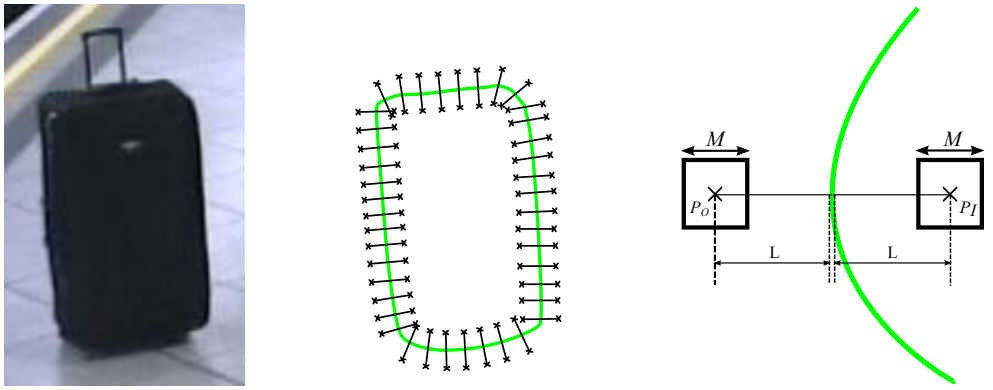


Figure 10. Proposed scheme for abandoned and stolen object discrimination.

The proposed approach is based on the spatial boundary contrast metric. For each pixel of the contour, segments normal to the contour's curve, are defined. The values of the pixels on both ends of the segment are then compared. This comparison is performed by defining a small window centered in those pixels. This scheme is illustrated in Figure 11.
We have evaluated the proposed approach using the ASODds dataset [10]. In particular, we have used the real data that consists on foreground masks representing the stationary objects of the scene which contains three categories with increasing complexity. In addition, we have compared our proposal against three representative approaches based on edge, color and contour. The accuracy results indicate that the proposed approach achieves higher performance than the ED and CH approaches and slightly better than the CO approach. The use of real data (that is inaccurate) with varying complexity demonstrates the robustness of the approach.



(a) (b) (c)

Figure 11. Pixel color contrast detector: static foreground object (left), analyzed points along the boundary (center) and analyzed contour point (right)

2.2.2. Fall detection

Independent living tasks are sometimes related to computer vision, and most of the works in this field are focused on event detection. These events are detected in many ways: by placing cameras in the environment, by mounting cameras on the person of interest, etc. One of the most interesting events is the fall detection. This is one of the events that are more dangerous for people who need care. This is why many research efforts have been directed toward the detection of this kind of event.

Different scenarios have to be considered when identifying different kinds of falls: falls from walking or standing, falls from standing on supports (e.g., ladders), falls from sleeping or lying in the bed and falls from sitting on a chair. An example of this last kind of fall is shown in Figure 12. There are some common characteristics among these falls as well as significant different characteristics. It is also interesting to note that some characteristics of fall also exist in normal actions, e.g., a crouch also demonstrates a rapid downward motion.

Figure 12. A typical fall from Sitting on a chair (frames from a simulated fall sequence). Extracted from [11]

In our work presented in [12], we developed an algorithm based in three main stages: a background modelling, a foreground segmentation and a detector of human parts. The proposed approach is schematically shown in Figure 13.

Figure 13. Block diagram of the proposed system for a general video input.

The input is a video from a camera placed somewhere in the environment. The resolution of those cameras is enough for generating a robust background model and detecting the person position, but do not requires a high definition image.

The background model is generated in batch mode using T frames, defining T as the number of frames where no foreground is visible. Once the background model is generated, it is not modified again.

For the subsequent frames, the model is compared with each of them. Using different metrics and morphological operations, a robust and reliable foreground mask is generated. This mask can contain people and other moving objects. These objects must be removed from the analysis, thus obtaining a foreground mask containing just people. Many people detectors have been designed in the state-of-the-art to cope with this issue, so this task lies out from the objective of this work.

The input of the human part detection stage is the foreground mask and the result of this stage is shown in Figure 14. Firstly, bounding boxes from the blobs in the scene are detected. Secondly, a part based analysis is performed in order to detect the centroids of the head, the torso and the legs. Finally, a pose estimation line is fixed linking the centroids.

Figure 14. Example of the bounding boxes (red) and the pose estimation line (blue) for three different poses. The blue crosses on the line are the centroids of each part. Extracted from [12].

For the final decision, is necessary to understand that a fall is the result of a complete process, so a fall detection is the result of detecting each of the different states in which a person is during that fall. In order to perform the detection, a finite state machine is defined. The inputs are the distances between the centroids (*D1* and *D2*) and the angles of each of the defined segments between the centroids ($\Theta 1$ and $\Theta 2$). The logic of the finite state machine defines the following states: no-fallen, false alarm, falling, fallen-wake up, and fallen-immobile. The resulting finite state machine is shown in Figure 15.

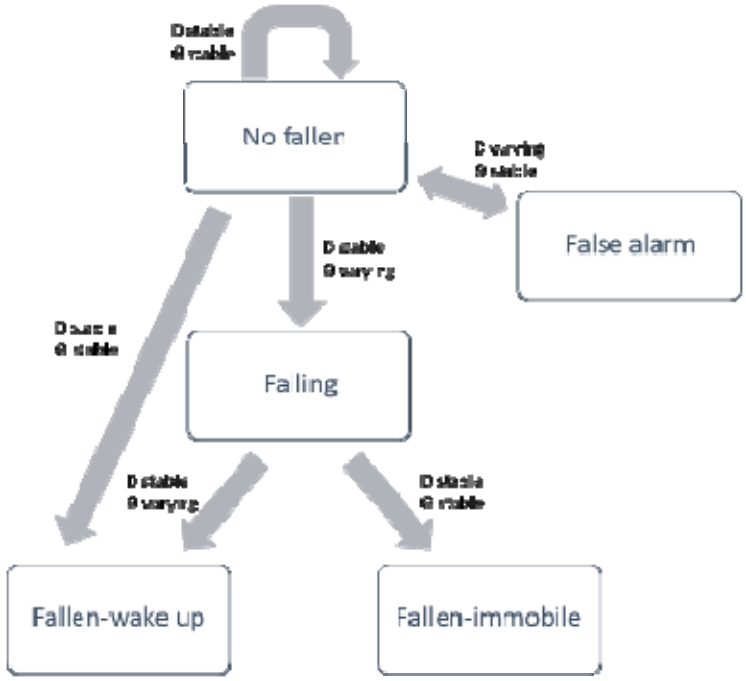


Figure 15. Finite state machine diagram. *D* and Θ are the inputs and the blocks are the states and the outputs (Moore like).

The final system is able to detect falls in medium complexity and low complexity scenarios. The computational cost is remarkably low and so is the required quality from the cameras. This results in an easily scalable and low cost system with average results in real scenarios.

2.2.3. Action recognition

The recognition of human-related events has recently become a relevant research area motivated by the variety of promising applications such as video surveillance, human–computer interaction and content-based indexing. Moreover, this interest can be also explained by the maturity of the employed low-level tools. Nevertheless, it still presents many challenges such as

the uncertainty of the low-level tools (e.g., object detection and tracking), the limited availability of training data, the similar appearance of different events and the modeling of complex relations.

In this area, we have explored the detection of human-related events that can be defined in a structured manner. The contributions developed in this research area are related with the structured event recognition (1 paper) and the use of contextual information to improve recognition which participated in the competition ICPR-HARL 2012 (1 master thesis).

2.2.3.1. A semantic-based probabilistic approach for real-time video event recognition

Many approaches have been proposed for event recognition which can be roughly classified into semantic and probabilistic. Semantic (or deterministic) approaches are based on defining rules to model the events. However, current approaches only describe a small portion of semantics (e.g., scene layout [2], event definitions [3]), they do not suggest the appropriate recognition strategies and they do not consider the uncertainty inherent to low-level observations and event definitions. On the other hand, the probabilistic approaches have shown a superior performance as compared to the semantic one. They accurately learn event models from training data achieving high precision within a domain and allowing an intrinsic uncertainty handling. However, they are not able to model complex relations and their usage is limited for different, albeit related, domains. In this situation, a combination of both approaches would be desirable for solving these limitations.

This work [10] addresses the above-mentioned limitations by introducing a new approach for event recognition that takes advantages of the accuracy of probabilistic approaches as well as the descriptive capabilities of semantic-based approaches. We start from Bayesian Networks (BNs) that are manually defined for real-time recognition of simple events. We propose a framework for complex event recognition based on hierarchical event descriptions that can be applied to a large variety of domains. The contribution of this work is three-fold. First, a state-of-art approach is integrated for event representation. The hierarchy of this representation model allows to apply recognition strategies suitable to each event type. Hence, a two-layer structure is defined for recognizing simple and complex events. Simple events are recognized by means of BNs, but in this work BNs are created automatically. The second extension regards the recognition of complex events by coupling the BNs with probabilistically-extended Petri Nets (PNs). The third extension defines a methodology to convert the event descriptions into their recognition models. Thus, BNs and PNs are built automatically from respectively, simple and complex event descriptions. We demonstrate the validity of the proposed approach for recognizing human–object interactions in the video monitoring domain. Experimental results show that it outperforms the widely used deterministic approach for recognizing events performed by different people in diverse scenarios whilst operating at real-time.

As a fundamental aspect of the proposed work, we use the Scene entity that represents the each domain by means of hierarchical descriptions of the scene objects (Object entity), their relations (Event entity) and additional information (SceneContext entity). We propose to exploit their relations (depicted in Figure 16) for achieving an effective recognition of events.

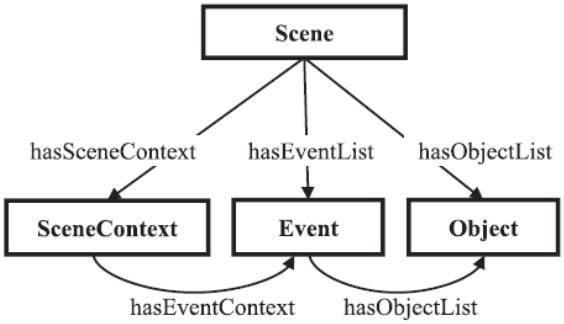


Figure 16. Entity relationships exploited for event recognition.

We design an event recognition framework composed of four modules as shown in Figure 17. The first module detects the objects of interest (i.e., the defined Object entities) from a video sequence. Then, the second module extracts the features required for event recognition. After that, a two-layer structure recognizes events considering the uncertainty of the analysis process being guided by the hierarchical event representation. First, the short-term layer performs the detection of simple events that are characterized by their occurrence in short-time periods. A BN is defined for each event based on its description. Then, the long-term layer recognizes the complex events that present a temporal relation among its counterparts. A probabilistically extended PN is defined for each hierarchical event representation composed of simple and complex events. Further details are available at [13]. The proposed combination addresses the limitations of the BN (not being able to model temporal event composition) and PN (deterministic detection) approaches. Note that this framework can fit the needs of a large variety of application domains by representing the prior knowledge and implementing the appropriate techniques for object detection and feature extraction.

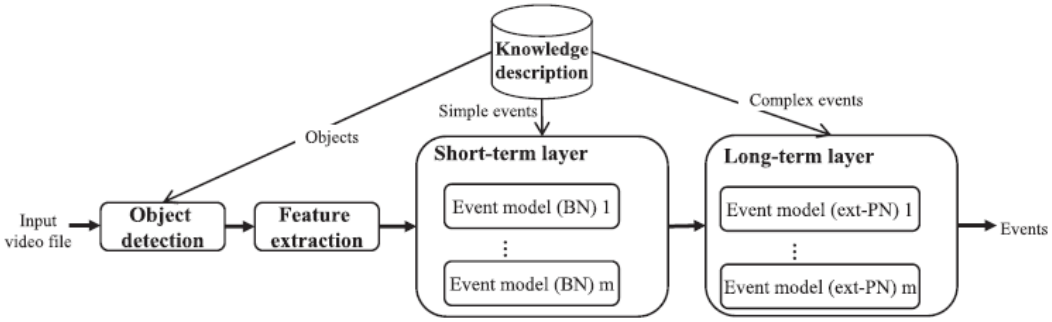


Figure 17. Proposed framework for the recognition of events.

Each event is recognized via a PetriNet manually defined as the example given in the following figure:

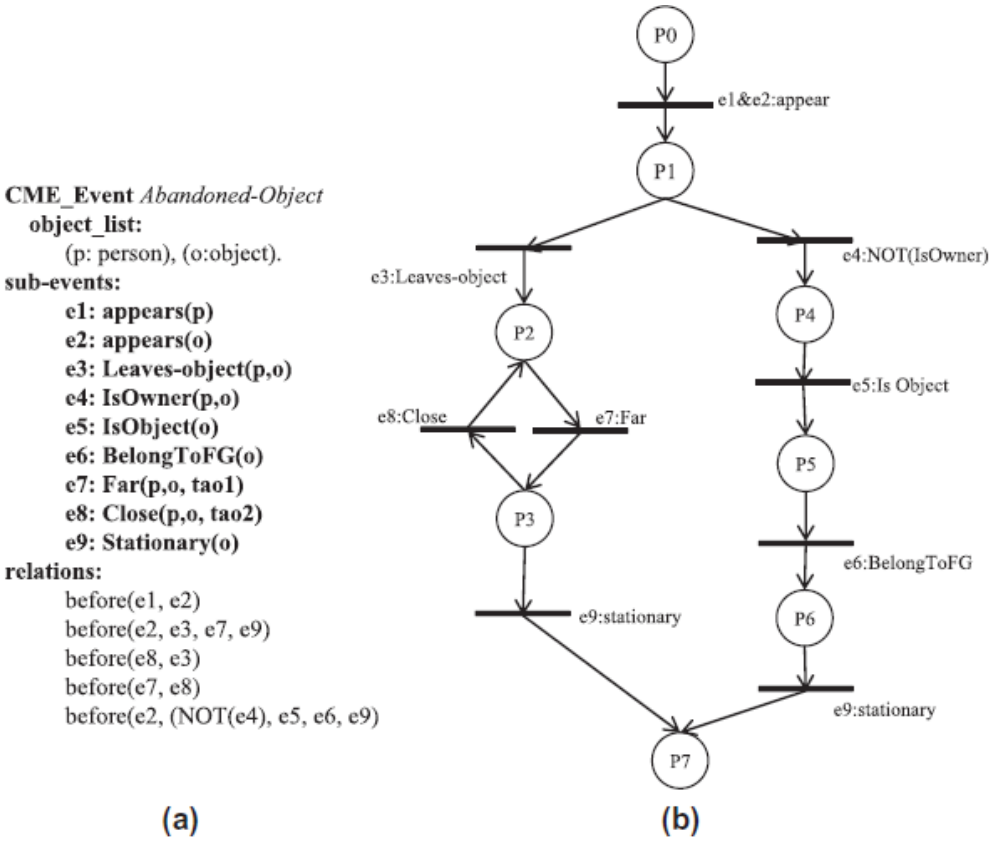


Figure 18. Complex event Abandoned-object modeled for the video surveillance domain. Data correspond to (a) semantic definition and (b) the corresponding PN.

The following figure shows event recognition examples for the different categories. It should be noted the increase in the number of false positives as we analyze more complex categories

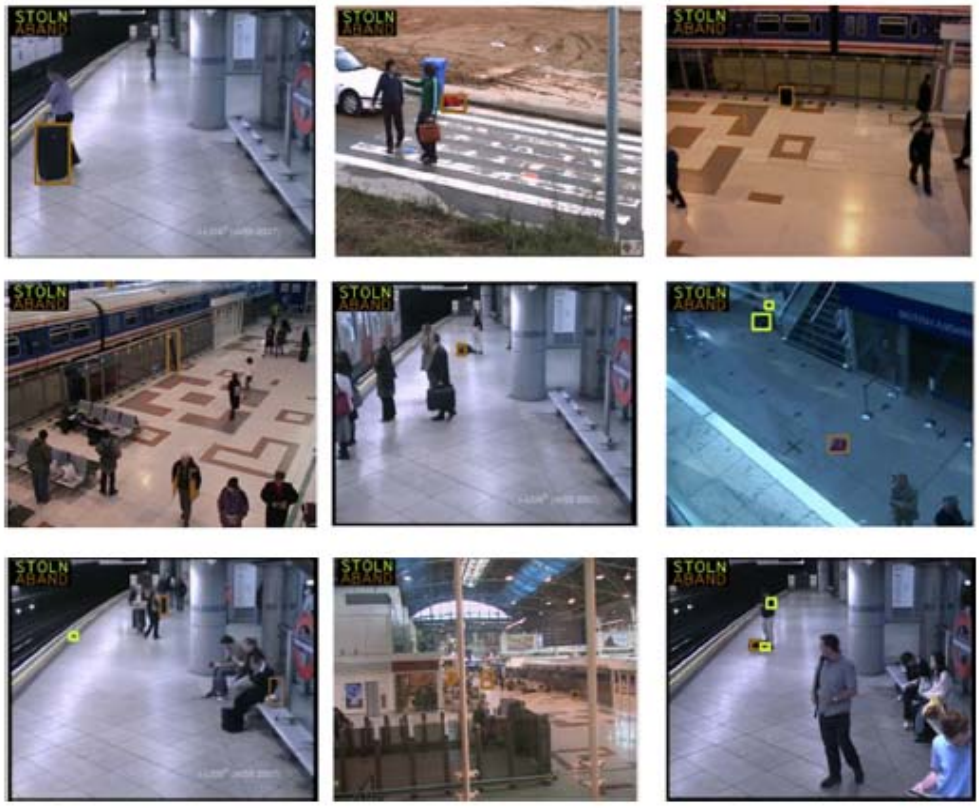


Figure 19. Event detection examples for uncontrolled environments. Rows 1, 2, 3 and 4 correspond to categories C1, C2, C3 and C4. (From top-left to bottom-right): CantataMultitelCam2_018 (frame 950), CantataMultitelCam1_013 (frame 1548), CantataMultitelCam1_013 (frame 1745), AVSS_AB_Easy (frame 2451), HERMES_Cam3_outdoor (frame 972), PETS06_S7_T6_B3 (frame 1641), PETS06_S5_T1_A4 (frame 2128), AVSS_AB_Medium (frame 2332), PETS07_S7 (frame 1755), AVSS07_hard (frame 3543), PETS06_S6_T3_H3 (frame 2329) and AVSS_AB_EVAL (frame 13430). The color codes correspond to the Abandoned-object (brown) and Stolen-object (yellow). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2.3.2. Analysis of interactions and activities in controlled environments

The approach developed within the EventVideo project is based on the event detection system that uses contextual information [13]. The VPULab approach contains the typical analysis stages (foreground segmentation, blob tracking, feature extraction and event recognition) and an additional one that considers contextual information that allows improving the event recognition rate. It detects 10 human-object and human-human interactions (all the events defined in the ICPR-HARL competition) based on features extracted from foreground blobs: blob velocity, blob trajectory, people likelihood, blob compactness, people skin and relative distances to contextual objects (tables, chairs, walls…). More details can be found at [14]. The following figure shows an example of the contextual information and depicts the block diagram of the approach.

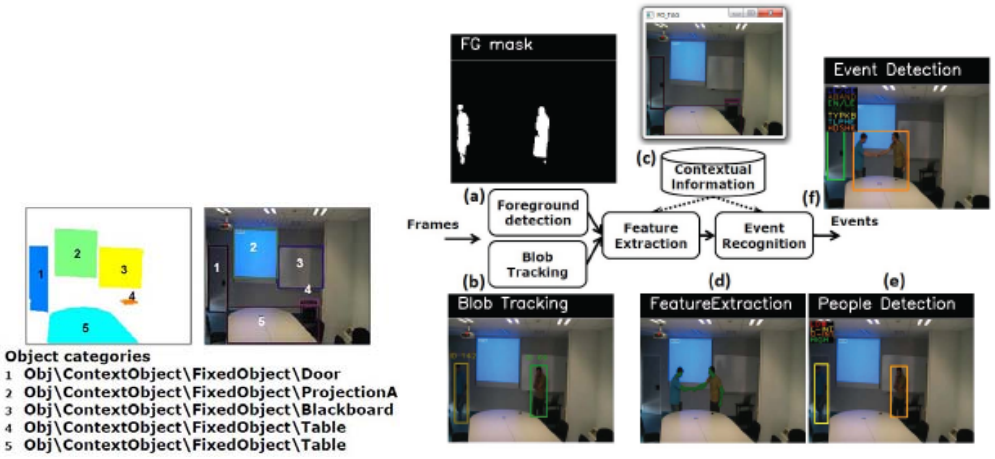


Figure 20. Example of contextual information used by the system (left) and Diagram of the VPULab event detection system (right).

The LIRIS dataset contains several human-object and human-human interactions in a controlled indoor settings captured with a static camera at a 720x576 resolution (25 fps). Each sequence contains 1-5 humans performing actions in short sequences (500-3000 frames). This dataset can be considered as a very realistic scenario as the events are performed in a natural way presenting several occlusions in most of the situations. Moreover, different viewing angles and distances to the camera were considered in the sequences. Besides colour and depth information were provided for the data composing two datasets: D1 (color+depth) and D2 (color). Sample frames are shown in Figure 21. More details can be found at http://liris.cnrs.fr/harl2012/

Here we present the results of the VPULab approach in the ICPR-HARL competition using the D2 dataset (only color) and a comparison with other participants. The comparison is given in the following table for the recognition task without spatial and temporal localization (e.g.,

bounding box and number of frames), that is, indicating if the event has been detected considering the entire sequence. It can be observed that the VPULab system presents an acceptable performance level compared to other participants. Although it obtained a low recall (36%), it has the second better value and it is the most precise system (66%). Compared to experiments in phase 1, we can observe that the VPULab approach has decreased its precision but increased the recall. Comparing the VPULab with other participants, we can observe that without using depth information, our system is able to achieve similar performance to the participants using D1 datasets (where foreground objects can be easily extracted based on depth data). Moreover, it can be observed that the two systems using key pose analysis present lower performance compared to the other approaches that do not use such technique. This can indicate that key pose estimation is a not sufficiently discriminative feature to differentiate events from other. Moreover, most of the participants used training data (from LIRIS-train) to detect in the test dataset. The low performance also indicated that pure machine learning methods are not suitable for event recognition as the variability in the executions of the same event is very high.

Equipo	Dataset	Recall	Precision	F-Score
ADSC-NUS-UIUC	D1	0.74	0.41	0.53
TATA-ISI	D1	0.08	0.17	0.11
VPULABUAM	**D2**	**0.36**	**0.66**	**0.46**
IACAS	D2	0.30	0.46	0.36

Table 1. Results of the ICPR-HARL 2012 competition (without localization). A description of the participant teams is available at http://liris.cnrs.fr/harl2012/

Figure 21 – Examples of human-related events of the LIRIS dataset (KEY. DI: Discussion, GI: Give Object. BO: Take Object. EN: Enter through a door. ET: Try to unlock a door. LO: Unlock a door. HS: Hand Shake. UB: Unattended Bag. KB: Keyboard typing. TE: Talking with telephone).

2.2.4. Feedback strategies for event detection

We present [15], a feedback-based approach to detect events in video surveillance. An estimation of the data complexity is used to adjust the computation effort of the analysis stages with the objective of improving the system performance (e.g. maintain accuracy while reducing computational cost).

Firstly, a video surveillance system (base system, see Figure 22) is developed with state-of-art techniques for the event detection task. The Foreground segmentation stage localize the moving objects based on adaptive background subtraction and statistical change detection. Additionally, the foreground data are filtered by a shadow removal stage based on the HSV color space. Subsequently, there is a blob extraction stage to group the connected regions followed by a blob tracking stage which, using a Kalman filter an spatial and color histogram distances, predicts the position for the blobs and computes their real position. Furthermore, the Blob classification stage determines the probability of each blob to be a person and the Feature extraction stage computes speed, direction and mean color of the blobs. With all the information above, stolen and abandoned object events are detected in the Event detection stage.

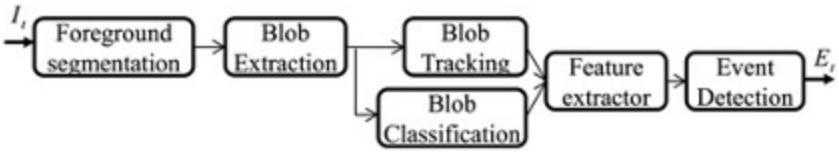


Figure 22. Block diagram of the base system for video event detection.

Secondly, a core structure is defined to support the feedback-based analysis. Such feedback aims to adjust the computational effort for different data complexities maintaining the accuracy of the analysis. This core feedback structure is based on two key ideas: availability of 'levels of detail for the analysis' (LoD) and the 'complexity estimation of the data' analyzed. A change in the LoD implies a variation on the computational effort and the accuracy of the performed analysis. It is assumed that an analysis with higher LoD will produce output results with higher or equal quality but never with less quality. In Figure 23 the proposed feedback structure is shown, where the 'actuator' decides, based on the estimated complexity of the data analyzed, which LoD is selected for the analysis of the processing stage.

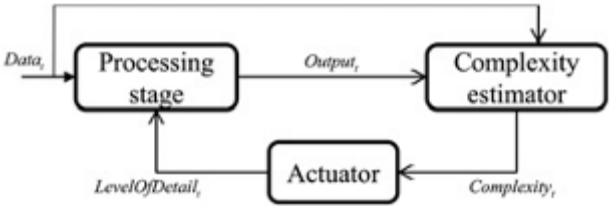


Figure 23. Proposed feedback structure for controlling the analysis effort.

Thirdly, the proposed feedback structure is introduced in the processing stages of the base system to improve the event detection accuracy and to adjust the computational cost of the stages to the complexity of the situation. For the Foreground segmentation, a 'ROI-based multiresolution analysis' has been implemented together with the segmentation algorithm as multiresolution analysis (pyramidal decomposition) is appropriate for obtaining different output accuracy. The LoD correspond to the levels in the pyramid structure (i.e. the analysis with different resolutions). As complexity estimator, we use the foreground percentage with respect to the image size trying to measure crowded situations that usually lead to environments more difficult to analyze. For the shadow removal task, the 'Maximization of the agreement between independent detectors' has been implemented. The well-known HSV shadow detection algorithm is decomposed into the intensity and chrominance parts, conforming two detectors which agreement is maximized. The LoD correspond to different agreement percentages between the detectors. As complexity estimator, we use the number of unknown pixels (no agreement) and the percentage of blobs correctly classified by the people detector. For the blob classification stage, 'Incremental focus of attention' is implemented, as the applied detectors only provide a score as output to indicate the likelihood of being people. Therefore each level of detail corresponds with the application of the available detectors (increasing the complexity of the detector with the LoD). As complexity estimator, we use the number of blobs detected as people or non-people understanding that a blob is classified as people (non-people) if its associated score is high (low) and the intermediate scores are penalized. For Event detection,

blobs are needed so they are extracted from the foreground segmentation, task that we can perform via several detectors, thus 'Incremental focus of attention' is a suitable approach to include. Three detectors are defined based on gradient, color and contour and they are sorted regarding their accuracy (the higher ranking, the better accuracy) and each LoD correspond to each detector. As complexity estimator, we use the number of correct detections into foreground or background for the detected blobs.

Finally, a system manager is included in the system to use the information generated by the feedback strategies for re-evaluating unknown events and adjusting the computational effort to the situation complexity (see Figure 24). It is in charge of estimating data complexity, selecting the level of detail for each processing stage and deciding the analysis strategy to apply.

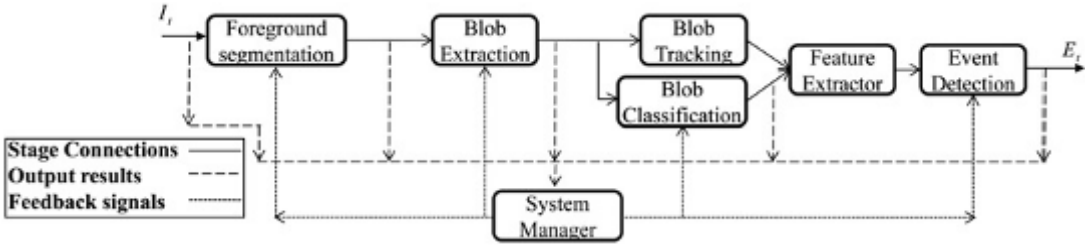


Figure 24. Extension of the base system to support feedback strategies.

We present a comparative evaluation of the proposed feedback-based event detection system and the base system in order to validate the utilization of feedback strategies. We obtain that the proposed feedback strategies improves the system accuracy because of the re-inspection of unknown events (initially discarded by the base system) with a higher LoD while the analysis with lower LoD almost maintain the performance. Additionally, the overall computational cost is decreased without reducing the system detection capabilities.

3. Conclusions and Future Work

In this work we have cover several event detection applications, going from undefined events (anomalies) to more defined events (abandoned/stolen objects, falls and interactions), and improving the state-of-the-art results in almost all of them by handling complex situations and using context information.

In one hand, the novel anomaly detection field has been studied in detail, and concrete improvements have been proposed. It is thought that improvements in this field could lead to improvements in several computer vision tasks.

In the other hand, the thoroughly researched field of action recognition has also been studied and improved. The stolen-abandoned task has been the one where the research work has been focused. Several and novel approaches have been proposed, and the state-of-the-art results have been overcome. Furthermore, feedback strategies to allow adaptation to different complexities have been proposed to handle long-term issues derived from the scene variation and to speedup applications.

More concrete task in the action recognition field has also been studied. A fall detection tool have been developed, and a newfangled algorithm to detect actions and interactions in controlled environments is proposed. The results and the idea of adding semantic information to the elements of a controlled environment presented of the latter proposal are state-of-the-art level.

In the future work, several research paths will be explored to handle the current challenges, which are mainly focus on long-term video-monitoring and, in the case of action recognition, in a shift of focus. Actually, a shift on the strategies pursued to perform the event detection task has been raised. It consist on changing from action recognition to recognition of objects functions and people intentions, as the same action could have a different meaning (e.g. imply danger or not) depending on the person who is executing the action, its intentions and the context involved.

References

[1] Caro, L.; SanMiguel, Juan C., "Anomaly Detection in Video Sequences, Caro Campos, Luis," Oct. 2013. Master thesis.

[2] P. M. Jodoin, V. Saligrama, and J. Konrad, "Behavior Subtraction". IEEE Transactions on Image Processing, vol. 21, no. 9, pp. 4244–4255, 2012. doi: 10.1109/TIP.2012.2199326.

[3] Ortego, D.; SanMiguel, J.C., "Stationary foreground detection for video-surveillance based on foreground and motion history images". IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 75-80, Aug. 2013. doi: 10.1109/AVSS.2013.6636619.

[4] Guler, S.; Silverstein, J.A.; Pushee, I.H., "Stationary objects in multiple object tracking". IEEE Conference on Advanced Video and Signal Based Surveillance. pp. 248-253, Sept. 2007. doi: 10.1109/AVSS.2007.4425318.

[5] Chang, J-Y; Liao, H-H; Chen, L-G, "Localized detection of abandoned luggage". EURASIP Journal on Advances in Signal Processing, article ID 11, pp. 1-10, Feb. 2010. doi: 10.1155/2010/675784.

[6] Bayona, A.; SanMiguel, J.C; Martínez, J.M, "Stationary foreground detection using background subtraction and temporal difference in video surveillance" IEEE International Conference on Image Processing (ICIP), pp. 4657–4660, Sept. 2010. doi: 10.1109/ICIP.2010.5650699.

[7] Ortego, D.; SanMiguel, J.C., "Multi-Feature Stationary Foreground Detection for Crowded Video-Surveillance". IEEE International Conference on Image Processing (ICIP), Paris (France), Oct. 2014.

[8] Zhou Wang; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P., "Image quality assessment: from error visibility to structural similarity". IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, Apr. 2004. doi: 10.1109/TIP.2003.819861.

[9] Porikli, F.; Ivanov, Y.; Haga, T., "Robust abandoned object detection using dual foregrounds". EURASIP Journal on Advances in Signal Processing, article ID 197875, 2008. doi: 10.1155/2008/197875.

[10] Caro, L., SanMiguel, J.C., and Martinez, J.M. : 'Discrimination of abandoned and stolen object based on active contours', Proc. of Int. Conf. on Advanced Video and Signal based Surveillance, Klagenfurt, Austria, 2011, Vol. 1, pp. 101–106. doi: 10.1109/AVSS.2011.6027302.

[11] Mubashir, M.; Shao, L.; Seed, L.; "A survey on fall detection: Principles and approaches". Neurocomputing. 2013. vol. 100, pp. 144-152. doi: 10.1016/j.neucom.2011.09.037.

[12] Cerro, S., Mártinez, J.M., "Detección de caídas para video-monitorización en entornos domésticos". May 2014. Diploma thesis.

[13] SanMiguel, Juan C. and Martinez, J.M., "A semantic-based probabilistic approach for real-time video event recognition". Computer Vision and Image Understanding, 116(9):937–952, Sept. 2012. doi: 10.1016/j.cviu.2012.04.005.

[14] Suja, S. "Análisis de interacciones y actividades en entornos controlados", Proyecto fin de Carrera, Ingeniería de Telecomunicación, Universidad Autónoma de Madrid, Dec. 2012.

[15] SanMiguel, J.C.; Martínez, J.M., "Use of feedback strategies in the detection of events for video surveillance". IET Computer Vision, vol. 5, no. 5, pp. 309-319, Sept. 2011. doi: 10.1049/iet-cvi.2010.0047.