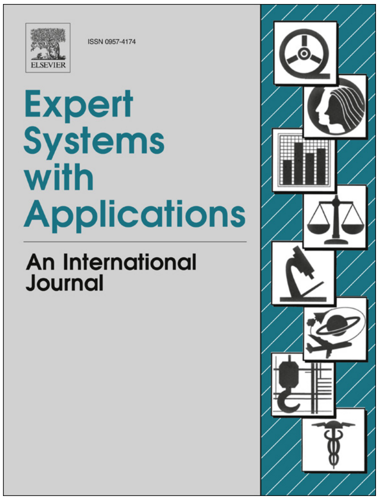


Expert Video-Surveillance System for Real-Time Detection of Suspicious Behaviors in Shopping Malls

Roberto Arroyo, J. Javier Yebes, Luis M. Bergasa, Iván G. Daza, Javier Almazán

PII: S0957-4174(15)00418-2
DOI: http://dx.doi.org/10.1016/j.eswa.2015.06.016
Reference: ESWA 10093

To appear in: *Expert Systems with Applications*

Received Date: 27 May 2014
Revised Date: 4 June 2015
Accepted Date: 13 June 2015

Please cite this article as: Arroyo, R., Javier Yebes, J., Bergasa, L.M., Daza, I.G., Almazán, J., Expert Video-Surveillance System for Real-Time Detection of Suspicious Behaviors in Shopping Malls, *Expert Systems with Applications* (2015), doi: http://dx.doi.org/10.1016/j.eswa.2015.06.016

Expert Video-Surveillance System for Real-Time Detection of Suspicious Behaviors in Shopping Malls

Roberto Arroyo, J. Javier Yebes, Luis M. Bergasa, Iván G. Daza and Javier Almazán.

University of Alcalá, Department of Electronics. 28871, Alcalá de Henares, Madrid, Spain.
E-mail addresses: roberto.arroyo@depeca.uah.es (corresponding author),
javier.yebes@depeca.uah.es, bergasa@depeca.uah.es, ivan.garciad@uah.es,
javier.almazan@depeca.uah.es

Abstract

Expert video-surveillance systems are a powerful tool applied in varied scenarios with the aim of automatizing the detection of different risk situations and helping human security officers to take appropriate decisions in order to enhance the protection of assets. In this paper, we propose a complete expert system focused on the real-time detection of potentially suspicious behaviors in shopping malls. Our video-surveillance methodology contributes several innovative proposals that compose a robust application which is able to efficiently track the trajectories of people and to discover questionable actions in a shop context. As a first step, our system applies an image segmentation to locate the foreground objects in scene. In this case, the most effective background subtraction algorithms of the state of the art are compared to find the most suitable for our expert video-surveillance application. After the segmentation stage, the detected blobs may represent full or partial people bodies, thus, we have implemented a novel blob fusion technique to group the partial blobs into the final human targets. Then, we contribute an innovative tracking algorithm which is not only based on people trajectories as the most part of state-of-the-art methods, but also on people appearance in occlusion situations. This tracking is carried out employing a new two-step method: 1) the detections-to-tracks association is solved by using Kalman filtering combined with an own-designed cost optimization for the Linear Sum Assignment Problem (LSAP); and 2) the occlusion management is based on SVM kernels to compute distances between appearance features such as GCH, LBP and HOG. The application of these three features for recognizing human appearance provides a great performance compared to other description techniques, because color, texture and gradient information are effectively combined to obtain a robust visual description of people. Finally, the resultant trajectories of people obtained in the tracking stage are processed by our expert video-surveillance system for analyzing human behaviors and identifying potential shopping mall alarm situations, as are shop entry or exit of people, suspicious behaviors such as loitering and unattended cash desk situations. With the aim of evaluating the performance of some of the main contributions of our proposal, we use the publicly available CAVIAR dataset for testing the proposed tracking method with a success near to 85% in occlusion situations. According to this performance, we corroborate in the presented results that the precision and efficiency of our tracking method is comparable and slightly superior to the most recent state-of-the-art works. Furthermore, the alarms given off by our application are evaluated on a naturalistic private dataset, where it is evidenced that our expert video-surveillance system can effectively detect suspicious behaviors with a low computational cost in a shopping mall context.

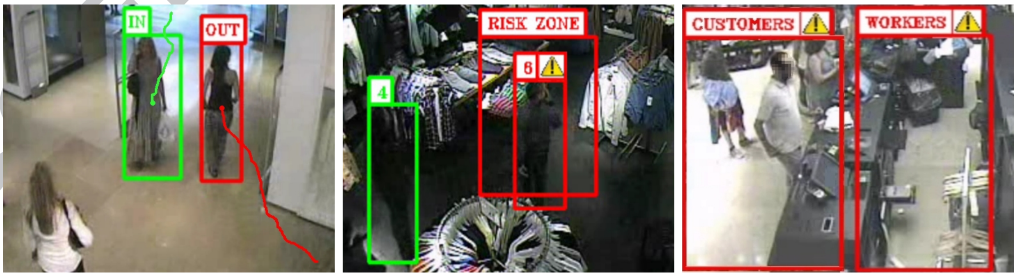
Keywords: Video-surveillance in shopping malls, Background subtraction, Human tracking, Occlusion management, Appearance features, Behavioral analysis

1. Introduction

In the current world, surveillance has become an essential element in a lot of daily activities to guarantee human security and property and assets protection. It is present in all kind of locations: banks, prisons, airports, parking lots, petrol stations, stores and any imaginable business or enterprise. Due to this, it exists an incipient need related to automating certain surveillance tasks for assisting security officers and allow them develop their work in a more efficient way.

Nowadays, the video images captured from cameras strategically located are the principal element in any surveillance system. For this reason, computer vision processing can be employed for extracting useful data from these videos and reasoning this information out with the aim of automating several video-surveillance tasks by giving off alarms when risk events are detected.

This paper is focused on a specific case of video-surveillance: to detect potentially suspicious human behaviors in shopping malls. In this scenario, there are some particular situations which must be analyzed, such as store entry or exit, loitering events that can culminate in a theft or situations where a cash desk is unattended, as shown in Fig. 1.

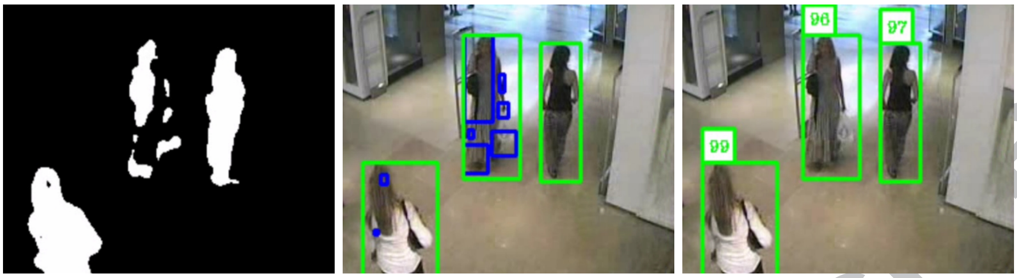


(a) Entry and exit.

(b) Loitering event.

(c) Unattended cash desk.

Figure 1: Alarms detected by our expert video-surveillance system in shopping malls.



(a) Background subtraction. (b) Blob fusion. (c) Tracking.

Figure 2: Video pre-processing and human tracking.

Before giving off the described alarms, there are some previous stages which must be considered in order to locate and follow the potentially suspicious people in the videos, as shown in Fig. 2. This previous process starts by making an image segmentation using background subtraction techniques. A comparison among the main background subtraction methods of the state of the art is carried out in this article with the aim of selecting the best suited for this specific application. In the next step of the process, the foreground objects, distinguished as blobs, are filtered by size and positional factors and grouped by our blob fusion algorithm. Afterwards, the tracking stage is essential for making an effective operation in the subsequent high-level phases such as trajectory analysis. Attending to this, an innovative tracking method is proposed, which consists in a two-step algorithm:

1. The detected objects to track (usually humans) are matched along the video sequence employing an optimization approach based on solving the association problem as a LSAP (Linear Sum Assignment Problem) (Easterfield, 1946) and considering the estimations of a Kalman filter (Kalman, 1960).
2. Occlusions between objects are managed applying a method based on visual appearance, in which several image descriptors are tested: GCH (Global

Color Histogram) (Novak & Shafer, 1992), LBP (Local Binary Pattern) (Ojala et al., 1994) and HOG (Histogram of Oriented Gradients) (Dalal & Triggs, 2005). The features extracted are compared through a metric based on SVM kernels similar to the proposed by Vapnik & Cortes (1995); Moghaddam & Yang (2000).

The usage of the introduced expert system increases the efficacy in a video-surveillance control center for a shopping mall in comparison with traditional methods, allowing human operators manage a higher number of cameras and their corresponding risk situations. Tests executed in naturalistic conditions demonstrate the reliability of our expert video-surveillance application and its efficient performance in real-time.

The main contributions of this paper are focused on the different innovative algorithms implemented in our expert video-surveillance system. First of all, although our segmentation stage is based on previous background subtraction techniques, we have carried out a complete study to choose the most proper for our system and, besides, we have designed a new blob fusion technique that reduces typical segmentation errors. In addition, our novel tracking algorithm represents an interesting new approach with respect to the classic state-of-the-art methods which were mainly based on trajectories, because we also use visual appearance information in occlusion situations for improving the performance of our proposal compared to other related works. Finally, the alarms analyzed by our system in shopping malls are a key contribution of this paper. To the best of our knowledge, this is the first work in the state of the art that considers the specific suspicious behaviors in shops processed by our expert application. We also contribute results and comparisons that corroborate the remarkable performance of our proposal.

The rest of the paper is organized as follows: Related work is discussed in Section 2. Section 3 describes the proposed methods for video pre-processing and human tracking. The methodology implemented in our expert video-surveillance system for detecting suspicious behaviors in shopping malls and giving off the corresponding alarms is explained in Section 4. Experiments and discussion are presented in Section 5 and the conclusions in Section 6.

2. Related Work

Since the 1960s, video-surveillance systems have evolved in a parallel line to their automation grade and three generations can be clearly differentiated, as stated by Räty (2010): 1^{st}) (1960 - 1980) Analog CCTVs and low level of automation. 2^{nd}) (1980 - 2000) Digital CCTVs and computer vision processing. 3^{rd}) (2000 - nowadays) Semi-automated video-surveillance systems.

The third generation of video-surveillance systems has achieved a certain grade of automation which allows to detect some suspicious human behaviors and to give off the corresponding alarms. In most cases, these systems follow a similar pattern to define their actuation methodology based on several sequential steps, which mainly are the foreground objects detection (a), tracking (b) and behavioral analysis (c):

(a) Foreground objects detection is a field with a large number of related works and an advanced level of satisfactory results, but the algorithm selection for a specific situation is complicated due to their different performance depending on the context and the scenario. There are several background subtraction methods commonly employed in automated video-surveillance applications, as is exposed in Stefano et al. (2011); Brutzer et al. (2011), and their results have been greatly improved in the last years using varied techniques such as the formulated by Maddalena & Petrosino (2008); Reddy et al. (2011); Verdant et al. (2011). On the other hand, some classic segmentation algorithms continue being widely applied, as the ones derived from Gaussian mixture models (Grimson & Stauffer, 1999; Zivkovic, 2004; Baf et al., 2008a), fuzzy models (Zhang & Xu, 2006; Baf et al., 2008b) or a combination of both (Baf et al., 2008c). Furthermore, other approximations can be interesting, such as the multi-layer background subtraction based on color and texture suggested by Odobez & Yao (2007). In any case, background subtraction techniques have successfully evolved and are robust enough to be applied in the proposed system. For this reason, a comparative study between the most representative enumerated algorithms in the shopping malls video-surveillance context is included in Section 3 for choosing the best option for our expert system.

(b) The bibliography about objects tracking is also extensive, but the results obtained in this field are not always satisfactory because to develop generic algorithms in this line is difficult. There are several tracking solutions for video-surveillance, and their applicability depends on the scenario: traffic situations (Alvarez et al., 2014), airports (Besada et al., 2005), maritime surveillance (Szpak & Tapamo, 2011), sports events (Kayumbi et al., 2008), places in poor lighting conditions (Wong et al., 2009; Gade & Moeslund, 2014), crowded environments (Zhao & Nevatia, 2004; Wu & Nevatia, 2006; Li et al., 2008, 2009; Xing et al., 2009; Kuo et al., 2010; Kuo & Nevatia, 2011; Ali & Dailey, 2012; Chau et al., 2014a,b; Badie & Bremont, 2014; Walia & Kapoor, 2014; Guan & Huang, 2015; Zhang et al., 2015), and so forth. However, there is not any realistic approach in related works for conveniently solving the problem of human tracking in a real-time video-surveillance system for shopping malls. Besides, generic algorithms similar to the presented in Foresti et al. (2005); Shah et al. (2007) do not offer satisfactory solutions for facing our particular problem, because the surveillance conditions in a shopping mall are very particular and, frequently, the images to process have low quality. It is better to focus on the specific challenges than to search generic tracking solutions, since each place has different and restrictive parameters. Attending to this, in this paper a new tracking algorithm is presented to fit into this kind of surveillance model for shopping malls. In Section 5, the theoretical contributions of our tracking approach will be highlighted in several experiments and they will be compared against the most recent state-of-the-art works.

(c) The behavioral analysis is an extensive field of study and, at present, some of the major research efforts inside the expert video-surveillance systems are directed in this line. There are multiple research topics related to this field: human gesture and posture estimation (Cristani et al., 2012), biometrics evaluation (Bashir et al., 2008), semantic-based video retrieval (Hu et al., 2004) or multi-camera behavioral analysis (Micheloni et al., 2005; Wang, 2013). In this work, we propose several innovative behavioral analysis algorithms for detecting risk situations in shopping malls which are mainly focused on alerting about customers entries or exits in the establishments, suspicious behaviors as loitering and unattended cash desk situations, as detailed in Section 4. This model implies a multi-camera system, in which assessing for every situation requires a different cameras placement.

Regardless of all the advances which has involved the third generation of video-surveillance, there are some persistent high-level problems which block a higher grade of automation (Pavlidis et al., 2001): the low cooperation between security systems, the extremely high valued assets insufficiently protected by obsolete technology or the excessive dependence on the intensive human concentration to detect and assess threats. Furthermore, there are other low-level problems which are not efficiently solved (Räty, 2010): the tracking in low-quality videos, occlusions between objects, algorithms executed in real-time, etc.

Due to the previous appreciations, some new concepts and approaches are appearing in order to increase the quality of automation in video-surveillance, included in what might be termed as a fourth generation of this kind of systems. The objective of this incipient generation is to search for new solutions, mainly based on networked, intelligent, multi-camera systems with integrated situational awareness of complex and dynamic scenes. While in the third generation the great efforts have been oriented in researching the issues underlying the above capabilities, the goal of the fourth generation is now on the applicability of these concepts in integrated systems, producing automated solutions for naturalistic surveillance problems (Xu, 2007). The application of these video-surveillance strategies provides reliable solutions which give safety personnel a higher level of situational awareness and allow they can react more quickly. Recently, these novel concepts have been applied in other similar expert video-surveillance applications for solving some of the exposed problems, such as the systems implemented by Castro et al. (2011); Albusac et al. (2011); Lim et al. (2014).

The work proposed in the present paper is framed in this fourth generation context with the aim of designing an innovative system and fomenting the evolution and automation of video-surveillance systems and their orientation to naturalistic situations.

3. Video pre-processing and human tracking

The design of a robust basis for identifying the objects of interest on the video scenes and evaluating their trajectories along the time is essential. The aim is to get a high hit rate in these previous tasks for making easier the subsequent detection of suspicious behaviors and alarms in the automated video-surveillance application for shopping malls presented in this paper. Therefore, the errors in the video pre-processing and human tracking stages dramatically affect the effectiveness of the final system. Due to this, these errors must be reduced as much as possible.

The initial video-surveillance phases defined in this work are the following: background subtraction, blob fusion, detections-to-tracks association based on Kalman filtering and a LSAP solution and, finally, an occlusion management based on visual appearance. The main goal of these algorithms is minimizing the error propagation between them and providing useful information about human trajectories to the final alarm processes for improving the general performance of the system.

3.1. Background subtraction overview

The advanced level reached in background subtraction is evident according to the state of the art, as shown in Section 2. In this work, some of the most powerful background subtraction methods formulated in the last years are considered to compare their characteristics and select the best one for our expert video-surveillance system in shopping malls.

The different techniques studied and tested in this paper are the following: self-organizing method through neural network by Maddalena & Petrosino (2008), Gaussian mixture model by Zivkovic (2004), fuzzy model by Zhang & Xu (2006), fuzzy mixture of Gaussian by Baf et al. (2008c) and multi-layer background subtraction based on color and texture by Odobez & Yao (2007). BGS library (Sobral, 2013) is used to implement the algorithms previously enumerated. This library contains open versions of the different background subtraction methods which are employed in this paper. Their parameters are adjusted by ROC computation on a training subset with the aim of maximizing detection rate and keeping a low false alarm rate.

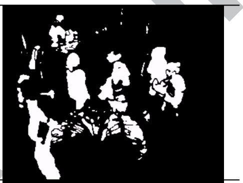
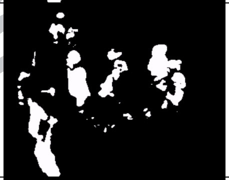
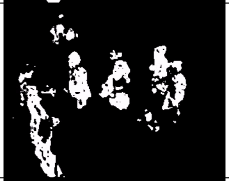
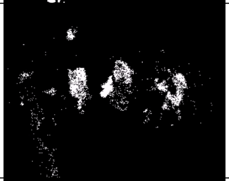
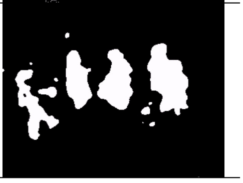
Table 1 presents a quantitative evaluation of every method, which includes their processing capacity in fps and the background subtraction metrics described in Izadi & Saeedi (2008): DR (Detection Rate), Spec (Specificity) and FAR (False Alarm Rate).

Table 1: Comparative study of the background subtraction statistics for each method adapted for the video-surveillance application in shopping malls context.

Method	Processing capacity	DR	Spec	FAR
Maddalena & Petrosino (2008)	33.27 fps	59.32 %	93.20 %	42.42 %
Zivkovic (2004)	62.45 fps	61.19 %	97.95 %	17.70 %
Zhang & Xu (2006)	19.73 fps	54.36 %	97.41 %	23.46 %
Baf et al. (2008c)	38.46 fps	22.14 %	99.62 %	9.92 %
Odobez & Yao (2007)	20.18 fps	82.03 %	98.69 %	15.30 %

In addition, Table 2 includes some visual segmentation examples for each method which had been taken from the employed private dataset described in Section 5, which contains shopping mall scenes with people crowd, low illumination, shadow effects, noise, etc., in order to visualize the algorithms performance in extreme (but naturalistic) conditions.

Table 2: Conclusions and visual comparison between the background subtraction methods adapted for the video-surveillance application in shopping malls context.

Method	Conclusions	Examples
Maddalena & Petrosino (2008)	In a shopping mall, people is constantly changing the place of clothes and coat stands. Due to this, background model must be updated frequently, and this algorithm is not adaptive enough for this situation.	
Zivkovic (2004)	Performance of this method has the finest quality-cost relationship. Although it can generate some noise and shadow effects, this technique provides acceptable results in a low processing time.	
Zhang & Xu (2006)	This algorithm is relatively consistent under pixelation and procures a reasonable effectiveness, but it has a high computational cost which can be a barrier in low-cost processing implementations.	
Baf et al. (2008c)	In spite of its low computational cost, pixelation and low illumination situations affect excessively to this method, which are very negative limitations for the proposed video-surveillance system.	
Odobez & Yao (2007)	The most effective of the algorithms: it defines clearly and completely human figures even in conditions of low illumination and reduces shadow effects. The inconvenient can be its elevated computational cost.	

Processed scene	Ground-truth

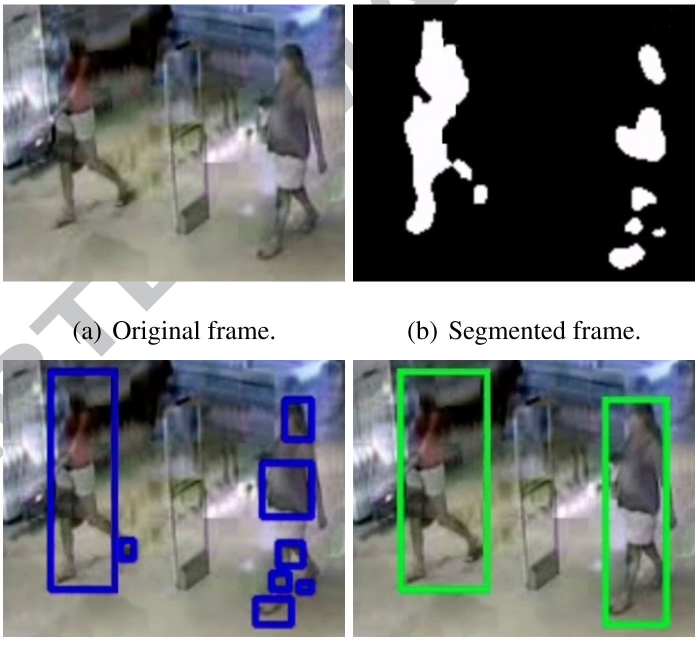
Attending to the evaluation, the most effective background subtraction approach for the surveillance scenario studied in this paper is the proposed by Odobez & Yao (2007). It yields high detection rates even in challenging conditions as low illumination or pixelation, which are very common in several low-cost video-camera pre-installed systems in shopping malls. However, the method suggested by Odobez & Yao (2007) is an expensive algorithm in computational cost terms, and its real-time performance in a multi-camera implementation only can be deployed in systems with a very high processing capacity, which is not our goal. For low computational cost solutions capable of working in real-time, the method proposed by Zivkovic (2004) is a reasonable alternative and it is selected for our expert video-surveillance application, because it also provides effective results (which are improved in the tracking stage) and allows managing more cameras in real-time.

As a final task for this stage, the segmented images obtained with background subtraction are filtered applying the dilation ($\oplus$) and erosion ($\ominus$) iterations described in Eq. 1, where S is the segmented image, F is the resulting filtered image and K is a $3x3$ kernel with the anchor at its center. The objective of this operation is, firstly, filling spaces in granulated foreground objects with dilation, after this, cleaning soft speckle noise with two erosion iterations and, finally, regularizing the silhouettes with a dilation.

$$F = ((S \oplus K) \ominus K \ominus K) \oplus K \tag{1}$$

3.2. Blob fusion

The designed blob fusion algorithm has the objective of correcting possible deficiencies in foreground objects detection produced by segmentation errors, as incomplete or fragmented person figures in the segmented images. In some situations, the video frames have excessive noise or data loss due to the video transmission from the shopping malls to the control center (see Section 4 and Fig. 6 to understand these problems derived from the data communication model and the usage of IP cameras). These difficulties cause poor segmentation results in some cases, as is presented in Fig. 3, where a frame with data loss and strong pixelation gives several foreground blobs where there is only one person, which is corrected with our blob fusion method. This approach is also supported by Brutzer et al. (2011), which opens a discussion about the usage of post-processing techniques for correcting background subtraction errors.



(a) Original frame. (b) Segmented frame.

(c) Without blob fusion. (d) With our blob fusion.

Figure 3: Blob fusion application in a frame with a video data loss which produces incomplete person figures after segmentation.

Our blob fusion technique consists of combining foreground blobs identified in a frame which can potentially correspond to one person. Let us define the set of detected blobs B, where each element b_i is characterized by the tuple (x_1,y_1,x_2,y_2). These values represent the coordinates of the upper left and bottom right corners of the detected rectangular blobs. In general, some elements of the set can correspond to parts of the same person, and the objective is to search certain position and size associations for fusing these elements and obtain the correct dimensions of each final detected object. In order to find these objects, three conditions must be fulfilled. Firstly, a pair of blobs are considered as fusionable candidates if the dimensions of any of them are lower than a limit l_1 (see Eq. 2). After this, it is necessary to look if these blobs are close to each other in the image vertical Y axis, considering a limit l_2 (see Eq. 3). Later, if previous conditions are satisfied for the pair of candidate blobs, another condition is analyzed for determining if they are in correlative positions in the horizontal X axis (see Eq. 4). Finally, if the pair of candidates also satisfies this last condition, they are fused, taking the coordinates for the new blob as defined in Eq. 5. As appreciation, when a new fused blob (b_{new}) is added to B, the blobs b_i and b_j which generated it are deleted from the set, where $i \neq j$. The described process is reiterated for each pair of elements belonging to B. The blob fusion method is repeated until n (the size of B set) do not change in a complete iteration for all the pairs.

Fusion condition 1:

$$\begin{gathered}((x_{2_{b_i}} - x_{1_{b_i}}) \cdot (y_{2_{b_i}} - y_{1_{b_i}}) < l_1) \;\; \vee ((x_{2_{b_j}} - x_{1_{b_j}}) \cdot (y_{2_{b_j}} - y_{1_{b_j}}) < l_1) \\ \text{,where } l_1 \approx \frac{\sum_{k=1}^{n} (x_{2_{b_k}} - x_{1_{b_k}}) \cdot (y_{2_{b_k}} - y_{1_{b_k}})}{n}\end{gathered} \tag{2}$$

Fusion condition 2:

$$\begin{gathered}(y1_{b_j} < y2_{b_i} + l_2) \lor (y1_{b_i} < y2_{b_j} + l_2) \\ \text{,where } l_2 \approx \sqrt{\frac{\sum_{k=1}^{n}(y2_{b_k} - y1_{b_k})}{n}}\end{gathered} \tag{3}$$

Fusion condition 3:

$$((x1_{b_i} \geq x1_{b_j}) \land (x1_{b_i} \leq x2_{b_j})) \ \lor ((x2_{b_i} \geq x1_{b_j}) \land (x2_{b_i} \leq x2_{b_j})) \tag{4}$$

Fused blob:

$$\begin{aligned} x1_{b_{new}} &= min(x1_{b_i}, x2_{b_i}, x1_{b_j}, x2_{b_j}); \\ y1_{b_{new}} &= min(y1_{b_i}, y2_{b_i}, y1_{b_j}, y2_{b_j}); \\ x2_{b_{new}} &= max(x1_{b_i}, x2_{b_i}, x1_{b_j}, x2_{b_j}); \\ y2_{b_{new}} &= max(y1_{b_i}, y2_{b_i}, y1_{b_j}, y2_{b_j}); \end{aligned} \tag{5}$$

3.3. LSAP association of tracked objects

When the set of detected objects B is completely filtered, is the moment of associating its elements with the recognized ones in the previous frames. For the object tracking, an association method based on solving this problem as a Linear Sum Assignment Problem (LSAP) is proposed in this paper. The theoretical classical definition of this problem given by Easterfield (1946) is well-known, but its solution depends on the context where is applied and, for this reason, an adapted technique is implemented for the video-surveillance system proposed in this work.

Furthermore, Kalman filtering (Kalman, 1960) is employed to predict the future positions and sizes of the tracked objects based on a constant velocity model.

As a first approximation, a cost matrix C is constructed to optimize the associations between the objects in the previous frames and the found ones in the current frame, given by the set B and filtered in the previous system stage. Each element c_{ij} is obtained from a cost function (f^k) as expressed in Eq. 6.

$$c_{ij}(k) = f^k = \sqrt{|x_{0_{b_j(k)}} - x_{0_{t_i(k-1)}}|^2 + |y_{0_{b_j(k)}} - y_{0_{t_i(k-1)}}|^2} \tag{6}$$

Let us define some of the new elements in Eq. 6: $t_i \in T$ is interpreted as a set which contains the predicted positions from the Kalman filter and all the trajectory information of the objects tracked; k is considered as a discrete time variable, where $k-1$ denotes the system state in the previous frame and k the current state; and the pair (x_0, y_0) represents the centroid of an object in one of the sets (B,T), where $x_0 = \frac{x_1+x_2}{2}$ and $y_0 = \frac{y_1+y_2}{2}$.

On the other hand, a result matrix R is defined, where each element r_{ij} is calculated following Eq. 7.

$$r_{ij}(k) = \begin{cases} 1 & \text{if } \; t_i(k-1) \; \text{ is assigned to } \; b_j(k), \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

With C and R, the LSAP solution designed is completely characterized, and the optimization problem can be faced by iterating Eq. 8 to find all the associations.

$$(i,j) = argmin(f^k(i,j)) \quad \Big/ \begin{array}{l} i \in \{B(k)\} \\ j \in \{T(k-1)\} \end{array} \tag{8}$$

Attending to Eq. 6, 7 and 8, also considering the objects entrances and exits from the scene and the occlusions between them, our complete solution provided for the object tracking and association is represented by the pseudo-code defined in Algorithm 1.

As a first step, R is calculated to solve LSAP, obtaining the minimal cost combinations between the elements of $T(k-1)$ and $B(k)$. Afterwards, the previous information stored in $T(k-1)$ for each object is refreshed with its current information, given by its related $b_j(k)$ element in R. In the next stage, $T(k)$ is completed by adding the new objects appeared on the scene and deleting the disappeared ones. The new objects to track are identified when a $b_j(k)$ has not any $t_i(k-1)$ associated in R, and the disappeared ones when a $t_i(k-1)$ has not any $b_j(k)$ associated in R. As appreciation, the objects are not directly added or deleted from $T(k)$, they have an error margin time established with the aim of avoiding errors produced by temporal false positive or false negative objects. This e_{margin} is defined as the number of frames which an object must stay on the scene to add it permanently to $T(k)$ or to delete it in the inverse situation. The value assigned to e_{margin} is approximately equal to the frame rate of the processed video and is reduced to the half if the object has appeared or disappeared near the image limits.

The final step to complete the proposed method is to deal with one of the most important problems associated to any tracking algorithm: the occlusions between objects.

3.4. Occlusion management based on visual appearance features

When a tracked object is not occluded by others, it is easy to recognize it in the next frame, because its position has not changed excessively and can be predicted with techniques as the proposed tracking method based on Kalman pre-

Algorithm 1 LSAP association of tracked objects.

Input: B in k instant and T in $k-1$ instant.

Output: T in k instant.

Initialization: C is filled following Eq. 6, and R has initially all its elements as 0. Furthermore, m is considered as the size of $T(k-1)$ and n as the size of $B(k)$.

Algorithm:

```
{Stage 1. Solve R matrix by iterating Eq. 8.}
$solved = 0$;
while $solved < min\{m, n\}$ do
    $min_{c_{ij}} = +\infty$;
    for $i = 1$ to $m$ do
        for $j = 1$ to $n$ do
            if $c_{ij} < min_{c_{ij}}$ then
                $min_{c_{ij}} = c_{ij};\ min_i = i;\ min_j = j$;
            end if
        end for
    end for
    if row($min_i$) and col($min_j$) of R have all their elements as 0 then
        $r_{min_i min_j} = 1;\ solved = solved + 1$;
    end if
    $c_{min_i min_j} = +\infty$;
end while
{Stage 2. Associate the elements of $T(k-1)$ and $B(k)$ using R.}
for $i = 1$ to $m$ do
    for $j = 1$ to $n$ do
        if $r_{ij} == 1$ then
            $t_i(k)$ = $t_i(k-1)$ refreshed with the new tracking information provided by $b_j(k)$;
        end if
    end for
end for
{Stage 3. Update $T(k)$ with the appeared and disappeared objects.}
for $j = 1$ to $n$ do
    if col($j$) of R have all their elements as 0 then
        Add $b_j(k)$ in $t_{m+1}(k)$;
    end if
end for
for $i = 1$ to $m$ do
    if row($i$) of R have all their elements as 0 then
        Delete $t_i(k)$;
    end if
end for

{Stage 4. Manage $T(k)$ occlusions as explained in Section 3.4.}
```

dictions and LSAP association. However, if an object is in an occlusion situation, the methods which are only based on the object position or direction are inadequate, because, specially in long occlusions, the trajectory of an object can be substantially altered.

In a shopping mall context, tracking is focused on humans, which usually have irregular and abrupt movements while they are observing or trying on clothes or other products. For this reason, when an occlusion ends, the people can not be re-identified conveniently using predictions of their positions. In order to minimize the occlusion problems related to the non-regular trajectories of people in the studied environment, an occlusion management algorithm based on visual appearance is suggested in this paper. This method complements the LSAP solution proposed for the detections-to-tracks association, and the combination of both gives as a result a powerful and robust tracking system oriented to video-surveillance of human behaviors.

On the one hand, if the predicted positions for two or more objects in $T(k-1)$ correspond to only one object in $T(k)$, an occlusion has begun. On the other hand, if the predicted position for one object in $T(k-1)$ corresponds to two or more objects in $T(k)$, an occlusion has ended.

The mechanism for occlusion management proposed in this work starts when an occlusion is detected. At this moment, the visual features of everyone involved in the occlusion are encapsulated, and these occluded people are tracked as a single entity with the method described in Section 3.3, but the system knows during all the occlusion that this entity is composed by two or more individuals. Finally, when the occlusion finishes, the pre-occlusion features labeled with a numerical identifier for each individual are compared with the post-occlusion unlabeled features to relate them and re-identify each occluded person, thus, tracking them individually again. For a better understanding, Fig. 4 displays the described approach for a simple two-people occlusion, where the basic idea behind the occlusion management based on visual appearance can be perceived.

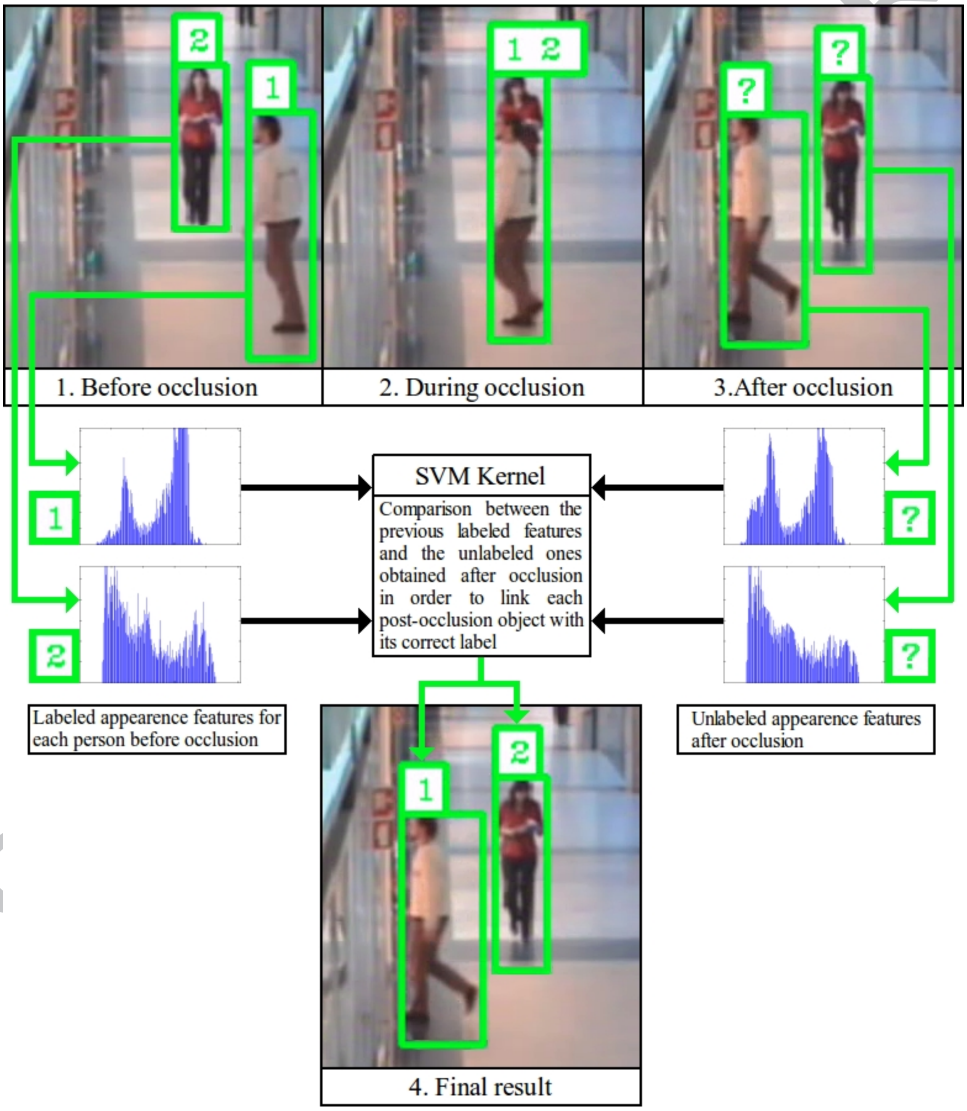

2
1
1 2
?
?
1. Before occlusion
2. During occlusion
3.After occlusion
1
2
SVM Kernel
Comparison between the previous labeled features and the unlabeled ones obtained after occlusion in order to link each post-occlusion object with its correct label
?
?
Labeled appearence features for each person before occlusion
Unlabeled appearence features after occlusion
1
2
4. Final result

Figure 4: Graphical description of the proposed occlusion management method based on visual appearance. The example shows a low complexity occlusion to understand the algorithm easily.

The example shown in Fig. 4 is a simplistic way to visually understand the proposed method. However, in more complex occlusions, there are other questions to solve related to the post-occlusion objects association, as can be seen in the complete diagram of the occlusion management presented in Fig. 5.

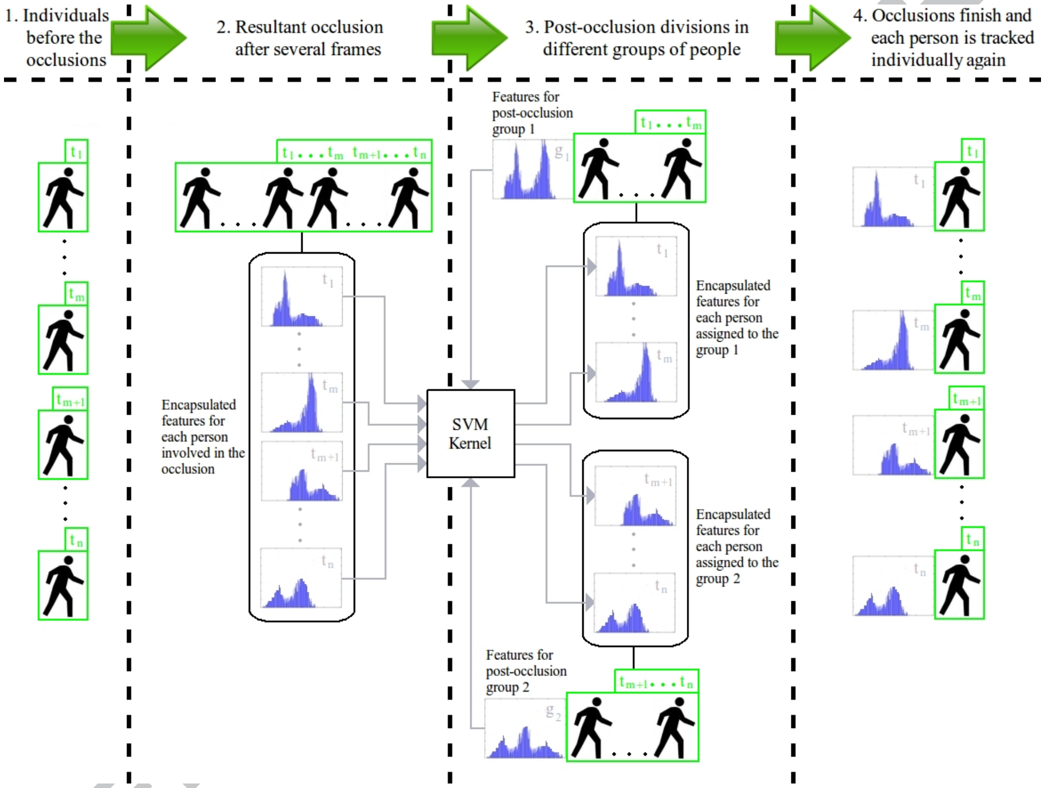

1. Individuals before the occlusions
2. Resultant occlusion after several frames
3. Post-occlusion divisions in different groups of people
4. Occlusions finish and each person is tracked individually again
t_1
t_m
t_{m+1}
t_n
$t_1 \dots t_m$ $t_{m+1} \dots t_n$
Encapsulated features for each person involved in the occlusion
SVM Kernel
Features for post-occlusion group 1
g_1
$t_1 \dots t_m$
Encapsulated features for each person assigned to the group 1
Encapsulated features for each person assigned to the group 2
Features for post-occlusion group 2
g_2
$t_{m+1} \dots t_n$

Figure 5: Diagram of the management mechanism for situations of multiple occlusions.

When an occlusion finishes, people not always leave the group individually, and it is common that some of them leave the occlusion in groups which continue occluded: for example, in occlusions of four people, they can leave it in two groups of two people. In these situations, the association process between the features saved for each person before the occlusion and the obtained ones from the groups after the occlusion, is not person-to-person (as in the example of Fig. 4). In this case, it is a person-to-group association, where the features of an individual before the occlusions are assigned to the group of people with the most similar characteristics according to the SVM kernel measure. These new groups of people originated after this separation continue being tracked as an occluded object. The features of every individual are kept on each group until the occlusion finishes completely or until the group goes out of the scene. If the occlusion finishes, each person involved is recognized and tracked individually again. Obviously, it must be remarked that the method described is focused on shopping malls where occlusions involve approximately 8 people as maximum, for very crowded groups of people, other research lines as the exposed in Zhao et al. (2012) can be followed.

The features employed for defining the human visual appearance are based on several image descriptors, which can be used individually or combined, because each applied descriptor has different properties which can be complementary. On the one hand, GCH (Novak & Shafer, 1992) is a low computational cost descriptor based on the color histogram of an object which, in spite of its simplicity, has a great performance in order to differentiate people by the colors of their clothes. On the other hand, LBP (Ojala et al., 1994) is a powerful texture descriptor also employed in human appearance with an efficient performance. Finally, HOG (Dalal & Triggs, 2005) describes the shape of a person through its gradient distribution, having a more expensive computational cost in comparison. The performance of all these descriptors for our occlusion management method is tested in Section 5, evaluating the results of each descriptor individually and combined between them.

With the aim of comparing the appearance features of a person before and after of an occlusion, some SVM kernel functions similar to the employed in Moghaddam & Yang (2000) are implemented: linear (Eq. 9), polynomial (Eq. 10), sigmoid

(Eq. 11) and radial basis (Eq. 12).

$$f_l(v_a, v_b) = v_a^T \cdot v_b \tag{9}$$

$$f_p(v_a, v_b) = (\gamma \cdot v_a^T \cdot v_b + \mu)^n \tag{10}$$

$$f_s(v_a, v_b) = tanh(\gamma \cdot v_a^T \cdot v_b + \mu) \tag{11}$$

$$f_r(v_a, v_b) = e^{-\gamma \cdot ||v_a - v_b||^2} \tag{12}$$

In kernels equations, v_a and v_b are the appearance features vectors of an object after and before the occlusion respectively, γ is a multiplier coefficient, μ is a coefficient of degree zero and n is the degree of the polynomial kernel. After evaluating the performance of these kernels in finding correspondences between features, the polynomial kernel is chosen, with $\gamma = 1$, $\mu = 0$ and $n = 2$. This kernel models the distances non-linearly, but a low grade (n=2) is chosen to prevent overfitting and higher computational costs. Besides, it compares not only individual features but combinations of them, remarking appearance similarities.

As a last observation, this appearance method might be also applied on the association of the non-occluded people along the sequence to fill the cost matrix values in the LSAP algorithm. However, if only appearance is used for tracking, the total computational cost will rise. Furthermore, in situations where there is not any occlusion, the results will be similar than employing a positional metric. For this reason, appearance is only used in situations where the positional reference of a person is lost, and these situations are mainly the occlusions.

4. Alarm detection in shopping malls

After defining a solid base for video pre-processing and human tracking, the alarm detection can be faced with a high guarantee of success in our expert video-surveillance system for shopping malls. In this context, the objective is to give off alarms which allow to detect certain risk situations or suspicious behaviors inside the stores. As shown in Fig. 6, the high-level view of the proposed system consists of a multi-camera approach where three differentiated shop zones are under monitoring. Each zone has a particular situation to evaluate: entry or exit of people in the entrance zone, loitering events in the shop interior and unattended cash desk situations in the payment area. The video information from the IP cameras in the shopping malls is sent employing a Web-based communication, and the data loss and pixelation effects derived from low bandwidth circumstances must be taken into account. The video data is processed depending on the camera situation and the alarms which must give off. Finally, the detected alerts are presented to the human operators in the control center, who can manage a higher number of cameras than using traditional video-surveillance systems.

4.1. Shop entry or exit control

The importance of alerting about the entries and exits of people in the shops resides in the detection of possible risky crowded situations when too many people enter inside the store and the vigilance on the exits of people showing suspicious behaviors like running away. The method proposed for this task is based on analyzing the trajectories of people in the entrance zones.

The first step consists in identifying the line of entrance to the shop because, depending on the particular store and the camera location, this can be placed in different positions and orientations. This task is manually made by human operators employing a simple interface in a previous system configuration, where the dividing line between the exterior and interior of the shop must be marked.

The alert process in this situation starts when a person over pass the defined

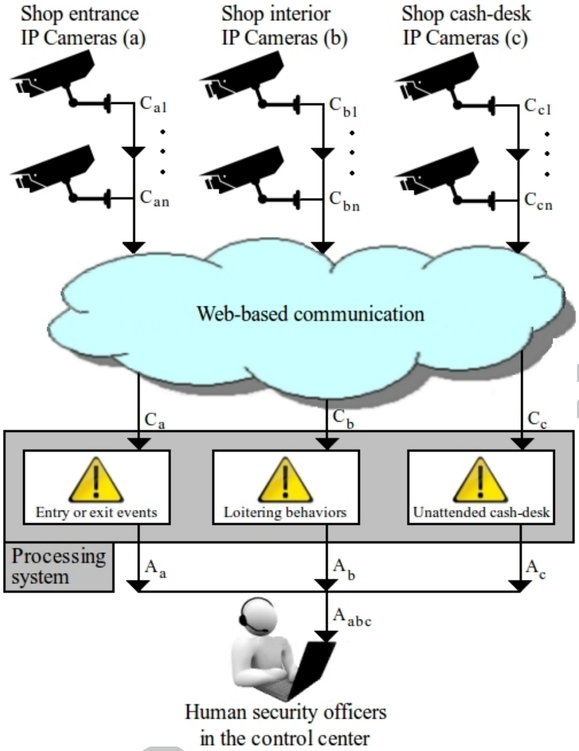


Figure 6: Multi-camera approach designed for the alarm detection defined in our expert video-surveillance system for shopping malls.

line r. If this person is considered as a tracked object in the previous video-surveillance stages and defined as $t_i(k)$, the alarm is given off if $r(k) \cap t_i(k)$. Afterwards, the next step must determine if the person is entering or exiting, which is evaluated considering its direction in the last frames with respect to the entrance line. The direction of a person is given in degrees and calculated following Eq. 13, which estimates the average angle of its trajectory in the last frames, considering the centroid position of the person in each moment as reference. The angular model employed can be observed in Fig. 7, where an example entrance line oriented on a $60°$ position is represented and the corresponding ranges of angles for entry and exit trajectories are shown. Finally, Fig. 8 depicts a real sample sequence of both entry and exit situations.

$$\overline{\alpha}_{t_i(k)} = \frac{\sum_{k=k_i}^{k_c} \left| arctan\left(\frac{|y_{0_{t_i(k)}} - y_{0_{t_i(k-1)}}|}{|x_{0_{t_i(k)}} - x_{0_{t_i(k-1)}}|} \right) + \beta_{t_i(k)} \right|}{k_c - k_i}$$

$$\text{,where } \beta_{t_i(k)} = \begin{cases} -90 & \text{if } ((x_{0_{t_i(k-1)}} > x_{0_{t_i(k)}}) \wedge \\ & \quad (y_{0_{t_i(k-1)}} > y_{0_{t_i(k)}})) \\ +90 & \text{if } ((x_{0_{t_i(k-1)}} > x_{0_{t_i(k)}}) \wedge \\ & \quad (y_{0_{t_i(k-1)}} < y_{0_{t_i(k)}})) \\ -270 & \text{if } ((x_{0_{t_i(k-1)}} < x_{0_{t_i(k)}}) \wedge \\ & \quad (y_{0_{t_i(k-1)}} < y_{0_{t_i(k)}})) \\ +270 & \text{if } ((x_{0_{t_i(k-1)}} < x_{0_{t_i(k)}}) \wedge \\ & \quad (y_{0_{t_i(k-1)}} > y_{0_{t_i(k)}})) \end{cases} \tag{13}$$

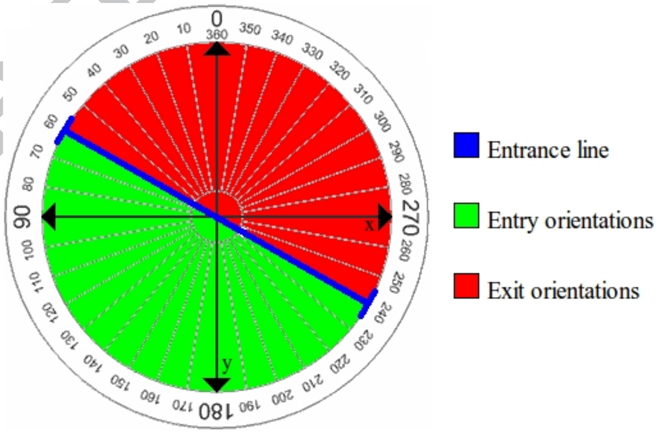


Figure 7: Example of the angular model proposed for the entry and exit trajectories directions if the entrance line is oriented on 60°.

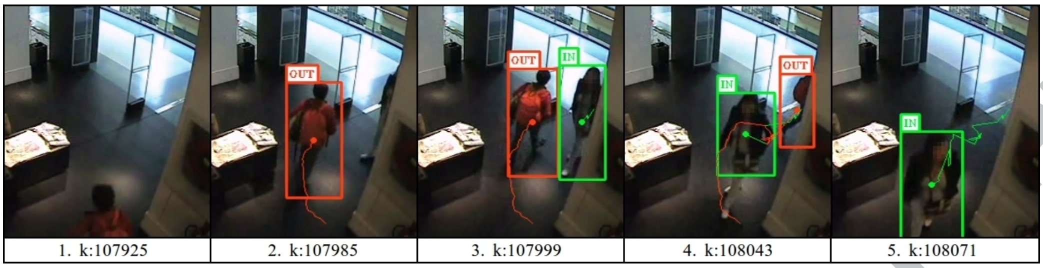


Figure 8: A real example of an alarm given off by the detection of entry and exit events: 1. A man appears on scene, but he is not still marked because he has not arrived to the entrance line. 2. When the man arrives to the entrance line, the system determines that he is exiting from shop. 3. A woman appears on scene with an entry direction. 4. After an occlusion produced by a crossing, the man and the woman bounding boxes preserve their trajectories. 5. The man disappears from the scene and the woman continues being tracked until she also goes out of the scene.

4.2. Detection of suspicious behaviors as loitering

In the interior zones of a shop, where the products are exposed, it is possible that malicious clients can commit a theft. In these situations, the shopping mall owners want to protect specific zones, specially where the more expensive articles are located. Attending to this, our system gives off alarms when a person has a suspicious behavior of loitering around a specific zone of the shop. As stated in Candamo et al. (2010), loitering is defined as the presence of an individual in an area for a period of time longer than a given time threshold.

The first step is to mark the specific risk zones, which can be selected by the human security officers with a simple implemented interface. After this, the system starts to evaluate the trajectories of the tracked people to determine if any of them is loitering in a risk zone for a long time. Our approach employs the positional tracking information of each object along the time and compares it with the position of the risk zone. If the person has its body completely inside the risk zone for an specific time (k_l), the loitering alarm is given off. Empirically, k_l is a time around 30 seconds that has been determined by security experts. This algorithm also considers the percentage of body that the suspicious person has inside the risk zone, increasing slowly its alarm ratio (k_a) if the person has not the whole body inside this zone, as is exposed in Eq. 14, where k_c is the current time, k_i is the time when the object enters inside the zone and w and h are the width and height of the bounding boxes of the suspicious person (t_i) and the risk zone (r).

$$k_{a_{t_i(k)}} = \sum_{k=k_i}^{k_c} \frac{w_{t_i(k)\cap r(k)} \cdot h_{t_i(k)\cap r(k)}}{w_{t_i(k)} \cdot h_{t_i(k)}} \tag{14}$$

The human security operators can easily visualize the evolution of people with loitering behaviors through the system interface, as shown in the real application example presented in Fig. 9. A color degradation between green and red is employed for representing the blob of a suspicious person depending of its k_a (see Eq. 14). When the loitering alarm is given off ($k_a \geq k_l$), the color of the blob is completely red and an alert icon appears for the suspicious person.

4.3. Unattended cash desk situations

The zone where the customers pay their purchases must be specially protected because the money contained in the cash register could be stolen. In order to alert the security officers about an unattended cash desk with someone loitering around it, an specific alarm is designed for preventing possible thefts.

The model proposed for this situation is analogue to the one exposed for the loitering alarms: if there is some person loitering near the cash desk as explained in Eq. 14 and there is not shop personnel detected in the cash desk zone, an alarm is given off, as was seen in the example C of Fig. 1. However, in this case the k_l estimated is around only 5 seconds, because this is a situation considered more risky and the control center must be quickly warned about it. In order to understand this kind of alarm, the three typical cash desk situations are shown in Fig. 10.

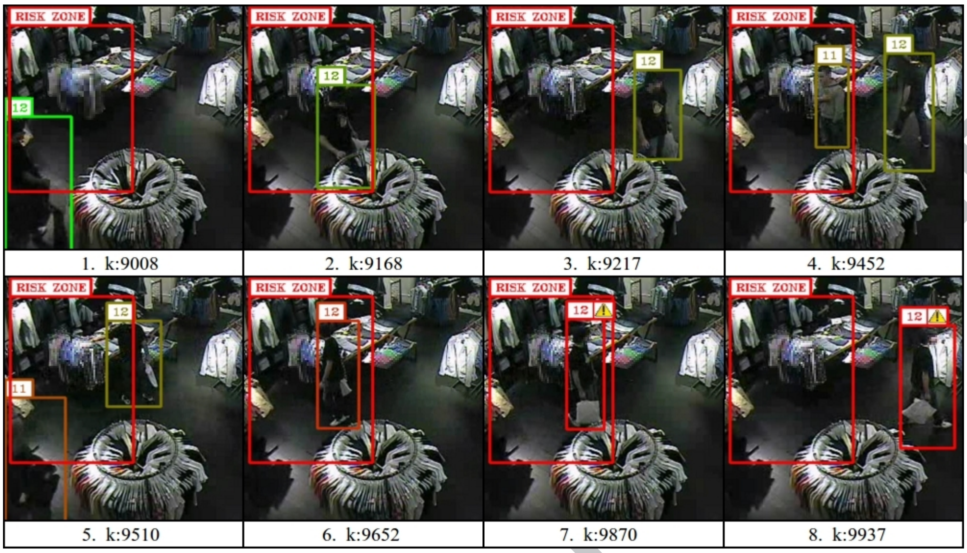
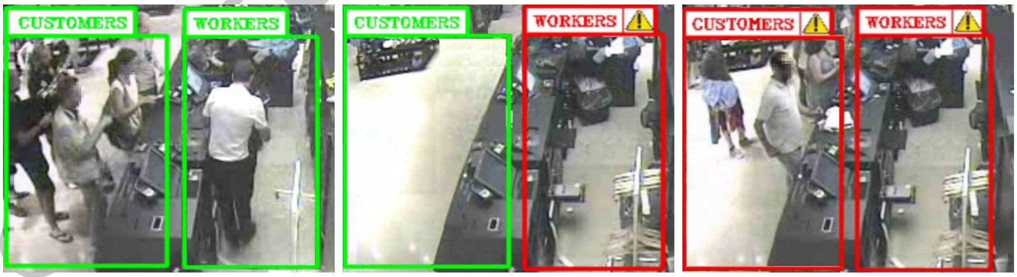


Figure 9: A real example of an alarm given off by the detection of a loitering event: 1. A person labeled as t_{12} appears on scene without alarm ratio. 2. t_{12} stays some time in the risk zone and its alarm ratio grows, as can be appreciated in its blob color. 3. t_{12} leaves the risk zone, but its alarm ratio is kept in case it enters again. 4. Another person labeled as t_{11} enters in the risk zone and, after staying some time, its alarm ratio grows. 5. t_{11} leaves the scene without giving off a loitering alarm because it has not stayed in the risk zone a time k_l. 6. t_{12} returns to the risk zone and its alarm ratio grows. 7. t_{12} produces a loitering alarm after exceeding k_l. 8. t_{12} leaves the risk zone, but it continues being tracked as suspicious in order to keep watch over it. **[Best viewed in color]**



(a) Attended cash desk.

(b) Unattended cash desk without people near it.

(c) Unattended cash desk with people near it.

Figure 10: A real example of the most typical cash desk situations.

5. Experiments

The main results derived from this work validate our expert video-surveillance alarm system for shopping malls. The experiments carried out are divided into two groups: the test of the proposed tracking algorithm and the analysis of the methods proposed for alarm detection in shops.

The experiments were carried out in a computer with the following features: an Intel Core i7 2.80 GHz processor and a 4 GB RAM. This is also applicable for the Table 1 presented in Section 3.1.

For tracking results evaluation, the public CAVIAR dataset (Fisher, 2004) is employed in order to compare our results with the obtained in other works where tracking is also used for multi-occlusion environments. CAVIAR is the public dataset most approximated to the conditions for this kind of video-surveillance application.

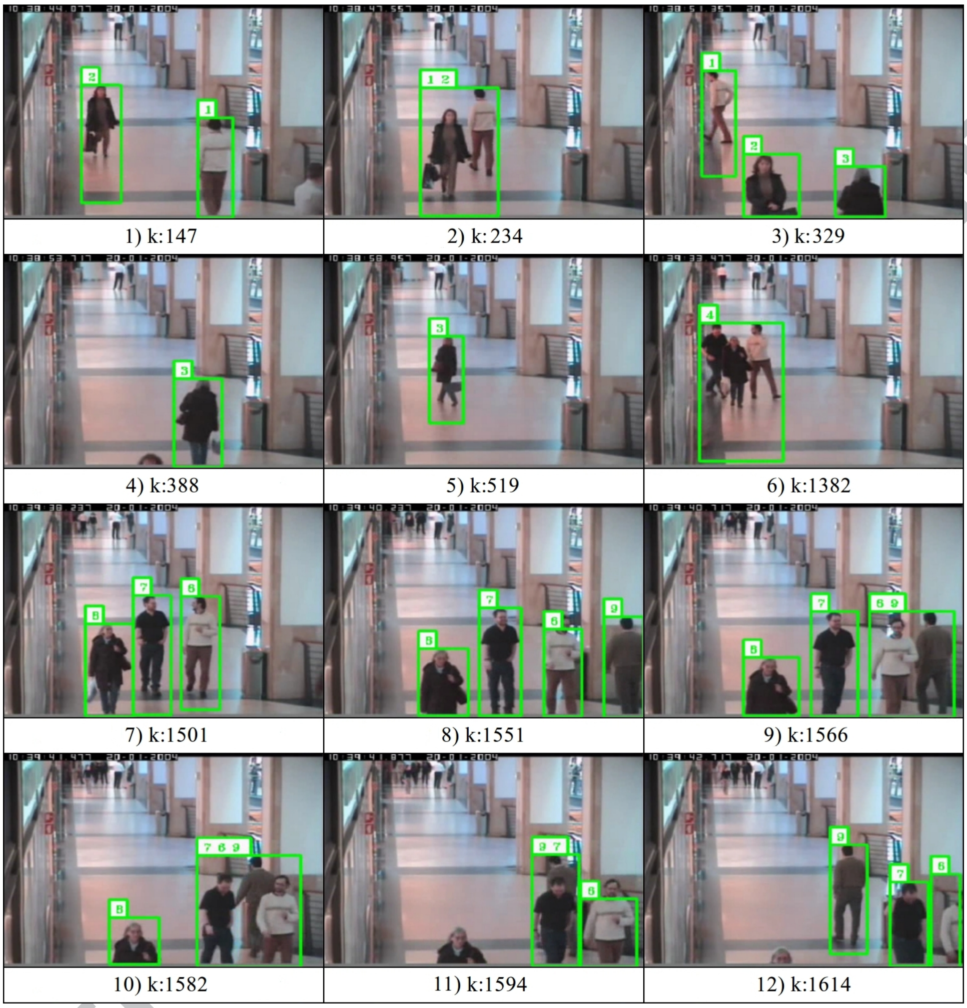
However, for the test of the alarm detection in shopping malls proposed in this work there is not any public dataset which contains videos with the specific alert situations defined in this paper. Due to this, a private dataset is employed, which is composed of videos with naturalistic shop situations and conditions as low illuminated zones, shadow effects, noise, frame loss and crowds. The original complete collection of videos from this dataset can not be published due to legal privacy restrictions, but their characteristics can be observed in Table 1 and Fig. 1, 2, 3, 4, 8, 9 and 10, where there are some frames presented from this dataset, which was digitized at 10 fps in the resolution of 352 x 288 pixels.

5.1. Tests for human tracking

The experiments proposed are focused on our people tracking approach and the occlusion management based on visual appearance.

With the aim of understanding the tracking evaluation carried out, Fig. 11 is presented. This sequence of images has been extracted from one of the videos of the CAVIAR dataset where the designed tracking method is tested. The occlusion management influence is obvious in this example and it can be seen how the people is correctly re-identified when the different occlusions finish. Furthermore, the behavior of other previous video-surveillance stages as background subtraction or blob fusion can be perceived. It demonstrates the importance of all the stages designed for our expert video-surveillance system and how they must interact to achieve the final goal.

There are several state-of-the-art papers where tracking for multi-occlusion environments is tested on the CAVIAR dataset, such as Zhao & Nevatia (2004); Wu & Nevatia (2006); Li et al. (2008). In these works, the tracking effectiveness is analyzed with some indicators related to the performance of the algorithms when occlusions between people appear on scene. In Zhao & Nevatia (2004); Wu & Nevatia (2006), the evaluation is focused on the occlusion duration: if occlusion is shorter than 50 frames, it is considered as a short-term occlusion (SO) and, otherwise, it is considered as a long-term occlusion (LO). In Li et al. (2008), it is also taken into account the number of people involved in the occlusion. Furthermore, there are differences between the events defined in Zhao & Nevatia (2004); Wu & Nevatia (2006) and the ones in Li et al. (2008), because the number of videos evaluated from the CAVIAR dataset and the criteria are slightly different in each one. In order to fairly test our tracking method, the criteria determined in Li et al. (2008) is followed. The results obtained are shown in Table 3, where our method is tested using the different visual appearance features defined for occlusion management in Section 3.4.



1) k:147 2) k:234 3) k:329

4) k:388 5) k:519 6) k:1382

7) k:1501 8) k:1551 9) k:1566

10) k:1582 11) k:1594 12) k:1614

Figure 11: Example of tracking evaluation carried out for the video “ShopAssistant1cor.mpg” from the public CAVIAR dataset: 1. t_1 and t_2 appear on scene. 2. The trajectories of t_1 and t_2 intersects and they get occluded. 3. The occlusion disappears and t_1 and t_2 are re-identified with its correct label. Furthermore, t_3 appears on scene. 4. t_1 and t_2 disappear from scene. 5. t_3 is abandoning the scene. 6. After several frames without activity, a group of three occluded people labeled as t_4 goes out from the shop. 7. The occlusion between the three people of the group labeled as t_4 is detected, and each person starts to be individually tracked as t_6, t_7 and t_8. 8. t_9 appears on scene. 9. An occlusion between t_6 and t_9 starts. 10. t_7 enters in the occlusion between t_6 and t_9. 11. t_6 abandons the occlusion and is correctly re-identified. 12. The occlusion between t_7 and t_9 finishes and both are re-identified with their correct labels. After this, the sequence finally concludes.

Table 3: Performance comparison between our approach for people tracking in the CAVIAR dataset and the proposed by Zhao & Nevatia (2004); Wu & Nevatia (2006); Li et al. (2008).

Method	**Occlusion complexity indexes**						**Overall**	**fps**
	2	**3**	**4**	**$\geq$5**	**SO**	**LO**		
Zhao & Nevatia (2004)	N/A	N/A	N/A	N/A	34/66	6/11	40/77 (51.9%)	≈ 3
Wu & Nevatia (2006)	N/A	N/A	N/A	N/A	39/66	9/11	48/77 (62.3%)	≈ 0.1
Li et al. (2008) (independent trackers)	12/18	10/17	2/6	1/4	12/17	13/28	25/45 (55.6%)	≈ 10
Li et al. (2008) (sequential tracking)	17/18	15/17	3/6	2/4	16/17	21/28	37/45 (82.2%)	≈ 10
Our method (with GCH)	15/18	10/17	0/6	0/4	11/17	14/28	25/45 (55.6%)	55.03
Our method (with LBP)	12/18	10/17	1/6	0/4	13/17	10/28	23/45 (51.1%)	53.76
Our method (with HOG)	13/18	12/17	1/6	1/4	13/17	14/28	27/45 (60.0%)	50.22
Our method (with GCH + LBP)	18/18	12/17	2/6	0/4	13/17	19/28	32/45 (71.1%)	52.85
Our method (with GCH + HOG)	17/18	14/17	1/6	1/4	15/17	18/28	33/45 (73.3%)	49.24
Our method (with LBP + HOG)	17/18	12/17	2/6	1/4	15/17	17/28	32/45 (71.1%)	48.32
Our method (with GCH + LBP + HOG)	18/18	16/17	3/6	1/4	16/17	22/28	**38/45 (84.4%)**	47.02

Attending to the tests of Table 3, the results of our method are in most cases superior than the ones achieved in Zhao & Nevatia (2004); Wu & Nevatia (2006) and in Li et al. (2008) with independent trackers. Moreover, for the combination of GCH + LBP + HOG appearance features, results are comparable to the obtained in Li et al. (2008) with sequential tracking, but the computational cost of our method is lower, having a processing capacity for tracking near to 50 fps. As can be appreciated in Table 3, the performance of our method is better if some image descriptors are combined for the occlusion management, because the features are complementary and reflect different characteristics of people. Hence, the hit rates obtained using GCH + LBP + HOG are slightly superior than the ones achieved in Li et al. (2008), which only uses one image descriptor for appearance: DCH (Dominant Color Histogram). However, for occlusions of more than 5 people, it also must be indicated that Li et al. (2008) slightly outperforms our method.

Currently, human tracking in crowded environments is a research topic in a constant evolution and, apart from the papers previously analyzed in Table 3, there are other more recent works where interesting tracking methods are applied, such as Xing et al. (2009); Li et al. (2009); Kuo et al. (2010); Kuo & Nevatia (2011); Ali & Dailey (2012); Chau et al. (2014a,b); Badie & Bremont (2014); Guan & Huang (2015); Zhang et al. (2015). In these recent works, tracking evaluation is not focused on occlusion situations, as seen in the results of Table 3. The applied methodology to determine the precision in these latest cases of study consists on evaluating the trajectory of the tracked people. According to these new metrics, MT (Mostly Tracked) means that more than 80% of the trajectory is correctly tracked, PT (Partially Tracked) means that between 20% and 80% of the trajectory is correctly tracked and ML (Mostly Lost) means that less than 20% of the trajectory is correctly tracked. GT is the number of trajectories in the ground-truth of the test videos. Depending on the work, a slightly different number of trajectories is evaluated as GT, but in our case we follow the criteria stated in Li et al. (2009) for the CAVIAR dataset, where 143 trajectories are evaluated. According to the previous considerations, we show more results in Table 4 using these metrics based on trajectories in order to compare the performance of our approach against the newest tracking algorithms appeared in the state of the art.

The results presented in Table 4 demonstrate how the performance of our method using appearance features based on GCH + LBP + HOG is comparable and slightly superior to the most recent state-of-the-art approaches. We achieve an 86.7% in MT, while the best state-of-the-art tracking methods recently published obtain a slightly lower precision of 86.4% (Chau et al., 2014a; Badie & Bremont, 2014). Besides, our algorithm also has a remarkable processing speed near to 50 fps, while the fastest method in the recent state of the art obtained a speed of approximately 31 fps (Zhang et al., 2015).

Table 4: Performance comparison between our approach for people tracking in the CAVIAR dataset and the proposed by Xing et al. (2009); Li et al. (2009); Kuo et al. (2010); Kuo & Nevatia (2011); Ali & Dailey (2012); Chau et al. (2014a,b); Badie & Bremont (2014); Guan & Huang (2015); Zhang et al. (2015).

Method	**GT**	**MT**	**PT**	**ML**	**fps**
Xing et al. (2009)	140	118/140 (84.3%)	17/140 (12.1%)	5/140 (3.6%)	≈ 5
Li et al. (2009)	143	121/143 (84.6%)	20/143 (14.0%)	2/143 (1.4%)	≈ 10
Kuo et al. (2010)	143	121/143 (84.6%)	21/143 (14.7%)	1/143 (0.7%)	≈ 4
Kuo & Nevatia (2011)	143	123/143 (86.0%)	19/143 (13.3%)	1/143 (0.7%)	≈ 7
Ali & Dailey (2012)	146	112/146 (76.7%)	34/146 (23.3%)	0/146 (0.0%)	≈ 3
Chau et al. (2014b)	140	120/140 (85.7%)	16/140 (11.4%)	4/140 (2.9%)	N/A
Chau et al. (2014a)	140	121/140 (86.4%)	15/140 (10.7%)	4/140 (2.9%)	N/A
Badie & Bremont (2014)	140	121/140 (86.4%)	12/140 (8.6%)	7/140 (5.0%)	≈ 5
Guan & Huang (2015)	146	119/146 (81.5%)	26/146 (17.8%)	1/146 (0.7%)	≈ 5.5
Zhang et al. (2015)	143	122/143 (85.3%)	19/143 (13.3%)	2/143 (1.4%)	≈ 31
Our method (with GCH + LBP + HOG)	143	**124/143 (86.7%)**	19/143 (13.3%)	0/143 (0.0%)	47.02

As deduced from the theoretical explanations given in previous sections, the successful performance of our method for people tracking is due to the usage of an association based on Kalman filtering and a LSAP optimization combined with an occlusion management that uses visual appearance based on features such as GCH, LBP and HOG. Other recent state-of-the art approaches such as Zhang et al. (2015) are only based on trajectories without using any kind of visual appearance information and, in consequence, they obtain worse results for tracking, as corroborated in Table 4. Due to this, other recent proposals that apply appearance descriptors such as Chau et al. (2014a,b); Badie & Bremont (2014) obtain results more similar to the achieved by our method, but they use features only based on color histograms and, for this reason, our algorithm has a slightly superior precision. Finally, in works such as Ali & Dailey (2012); Guan & Huang (2015), other techniques based on head tracking are carried out, but as shown in Table 4, these approaches do not perform a tracking so robust as the methods based in the whole body appearance.

5.2. Tests for alarm detection in shopping malls

The alarms designed in this paper for shopping malls are difficult to evaluate and compare because there is not any public dataset which completely fulfills the demanded requirements. Additionally, there is not any related bibliography available for some of the defined kind of alarms in similar scenarios. Fortunately, the private dataset of videos captured from cameras in naturalistic shopping mall situations which is employed in this work allows to carry out the necessary experiments to validate our system, also considering an associated ground-truth of alarms manually annotated.

For entrance and interior situations, this private dataset has two videos of one hour for each case, and for cash desk there is one video of one hour, five hours of real surveillance in total. The percentages of success obtained in the tests of these high-level alarms in shopping malls can be observed in Table 5 for each video and risk situation. Furthermore, it must be noted that for entries and exits there are 12 false negatives and 4 false positives (the remaining errors are entries marked as exits and vice versa), and for loitering there are only 2 false positives. In cash desk situations, there is only one alarm event that can be considered as risk situation, but the important conclusion for this kind of alert is that the system is stable and it does not generate false positives or negatives.

These results for alarm detection are satisfactory enough and, with these percentages of success, our expert video-surveillance system helps widely to the human security officers in the control center. By employing this application, a human operator can attend to the risk events of approximately the double number of cameras than by using a traditional monitoring system, because the additional information contributed by our automated system assumes an added value for the human security officer, specially when the fatigue starts concerning after several working hours. Besides, our method also can be employed for off-line batch processes of forensic video analysis, with the aim of evaluating legal matters in past records and highlighting people of interest and suspicious behaviors in the off-line videos.

Table 5: Results obtained for alarm detection in shopping malls employing a private naturalistic dataset.

	Alarms				
Video	**Entry**	**Exit**	**Loitering**	**Unattended c. d.**	**Overall**
Entrance A	426/480	435/512	0/0	0/0	861/992 (86.8%)
Entrance B	76/99	41/60	0/0	0/0	117/159 (73.6%)
Interior A	0/0	0/0	22/24	0/0	22/24 (91.6%)
Interior B	0/0	0/0	17/18	0/0	17/18 (94.4%)
Cash desk	0/0	0/0	0/0	1/1	1/1 (100.0%)
Overall	502/579 (86.7%)	476/572 (83.2%)	39/42 (92.8%)	1/1 (100.0%)	

6. Conclusions

Automated video-surveillance systems must employ effective low computational cost algorithms in order to process alarms in the greatest number of cameras possible with satisfactory and real-time results. There are a lot of works based on multi-camera systems which offer solutions with solid results but do not take into account the processing time and only can manage a reduced number of cameras in real-time, as is studied in Wang (2013). Our approach, focused on a complete surveillance of a shopping mall, can operate in a naturalistic multi-camera model efficiently, even in difficult conditions originated by video compression and low quality images derived from a previous data transmission to the control center. This efficiency allows managing several cameras for each risk situation considered: shop entrance cameras for entry and exit events, shop interior cameras for detecting suspicious behaviors as loitering and shop cash desk cameras for avoiding unattended cash desk situations. The study of this specific alarms in a shopping mall context by our expert video-surveillance system is a contribution to the state of the art, because there are not concrete related works that consider this specific kind of suspicious behaviors in shops.

The alarms given off by the system must be accurate and, due to this, the previous stages before processing the risk events produced in a shopping mall must be designed with the aim of reducing errors. In this paper, the tracking method based on a LSAP solution provides an efficient model in order to identify people in the video frames along the time, specially if its combination with the appearance algorithm is considered for managing occlusions in crowd situations. Nowadays, the tracking solution presented has an effectiveness superior or at least comparable to similar methods of the state of the art and our approach also procures a remarkable improvement in processing capacity. In the tracking results presented along this paper, the performance of our tracking method in occlusion situations has been compared to some classic algorithms such as Zhao & Nevatia (2004); Wu & Nevatia (2006); Li et al. (2008), where our method outperforms the performance of these approaches. The remarkable precision of our tracking algorithm is due to the addition of visual appearance information in occlusions management, because other recent state-of-the-art proposals such as Zhang et al. (2015) are only based on trajectories and they obtain worse results. However, in recent works such as Chau et al. (2014a,b); Badie & Bremont (2014) visual appearance is also used, but our tracking method obtains better results because our appearance is based on a combination of GCH, LBP and HOG features, while the previously cited works are mainly based on only color histograms information. In addition, there are other recent papers based on head tracking such as Ali & Dailey (2012); Guan & Huang (2015), but these approaches do not perform a tracking so robust as the methods based in the whole body appearance. For all these reasons, our tracking algorithm can be considered a great contribution to the expert and intelligent systems research, which could be also applied in other similar applications apart from the video-surveillance context described in this paper.

In future works, other type of features apart from the used in our people tracking could be proposed for describing visual appearance in this kind of systems, such as the fast binary descriptors, which have been satisfactorily tested in other computer vision areas related to video-surveillance such as place recognition (Arroyo et al., 2014a,b, 2015). As other future upgrade to our expert system, the conditions applied in blob fusion method can add a global score for each blob candidate in order to create a hypotheses search space in which to find the highest scoring fusion proposal. Moreover, other interesting research line would be the implementation of a similar system with collaborative cameras where 3D information can provide a higher situational awareness. In addition, the usage of stereo cameras instead of monocular vision could be an interesting upgrade in order to make easier the extraction of tridimensional data. Finally, some other interesting future research directions for the expert systems with applications community could be derived from the present work, such as obtaining a higher level of automation in the surveillance processes presented in this work. Currently, our system needs the supervision of a human operator to manage the alarms given off when suspicious behaviors are detected. However, powerful machine learning algorithms could be applied in the future with the aim of completely automatizing all the surveillance tasks.

Acknowledgments

This work has been supported with funds from the University of Alcalá through a FPI grant, the Spanish MINECO through the project Smart Driving Applications (TEC2012-37104) and the CAM through the project RoboCity2030 II (S2009/DPI-1559).

References

Albusac, J., Vallejo, D., Castro-Sanchez, J., & Jimenez-Linares, L. (2011). OCULUS surveillance system: Fuzzy on-line speed analysis from 2D images. *Expert Systems With Applications (ESWA)*, *38*, 12791–12806.

Ali, I., & Dailey, M. (2012). Multiple human tracking in high-density crowds. *Image and Vision Computing (IMAVIS)*, *30*, 966–977.

Alvarez, S., Llorca, D., & Sotelo, M. (2014). Hierarchical camera auto-calibration for traffic surveillance systems. *Expert Systems With Applications (ESWA)*, *41*, 1532–1542.

Arroyo, R., Alcantarilla, P. F., Bergasa, L. M., & Romera, E. (2015). Towards life-long visual localization using an efficient matching of binary sequences from images. In *IEEE International Conference on Robotics and Automation (ICRA)* (pp. 6328–6335).

Arroyo, R., Alcantarilla, P. F., Bergasa, L. M., Yebes, J. J., & Bronte, S. (2014a). Fast and effective visual place recognition using binary codes and disparity information. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 3089–3094).

Arroyo, R., Alcantarilla, P. F., Bergasa, L. M., Yebes, J. J., & Gamez, S. (2014b). Bidirectional loop closure detection on panoramas for visual navigation. In *IEEE Intelligent Vehicles Symposium (IV)* (pp. 1378–1383).

Badie, J., & Bremont, F. (2014). Global tracker: An online evaluation framework to improve tracking quality. In *IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)* (pp. 25–30).

Baf, F. E., Bouwmans, T., & Vachon, B. (2008a). Background modeling using mixture of gaussians for foreground detection - a survey. *Recent Patents on Computer Science (RPCS)*, *1*, 219–237.

Baf, F. E., Bouwmans, T., & Vachon, B. (2008b). Fuzzy integral for moving object detection. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1729–1736).

Baf, F. E., Bouwmans, T., & Vachon, B. (2008c). Type-2 fuzzy mixture of gaussians model: application to background modeling. In *International Symposium on Visual Computing (ISVC)* (pp. 772–781). volume 2.

Bashir, F. I., Usher, D., Casaverde, P., & Friedman, M. (2008). Video surveillance for biometrics: Long-range multi-biometric system. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)* (pp. 175–182).

Besada, J. A., Garcia, J., Portillo, J., Molina, J. M., Varona, A., & Gonzalez, G. (2005). Airport surface surveillance based on video images. *IEEE Transactions on Aerospace and Electronic Systems (TAES)*, *41*, 1075–1082.

Brutzer, S., Hoferlin, B., & Heidemann, G. (2011). Evaluation of background subtraction techniques for video surveillance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1937–1944).

Candamo, J., Shreve, M., Goldgof, D., Sapper, D., & Kasturi, R. (2010). Understanding transit scenes: a survey on human behavior-recognition algorithms. *IEEE Transactions on Intelligent Transportation Systems (ITS)*, *11*, 206–224.

Castro, J. L., Delgado, M., Medina, J., & Ruiz-Lozano, M. D. (2011). Intelligent surveillance system with integration of heterogeneous information for intrusion detection. *Expert Systems With Applications (ESWA)*, *38*, 11182–11192.

Chau, D., Bremond, F., & Thonnat, M. (2014a). Automatic tracker selection w.r.t. object detection performance. In *IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 870–876).

Chau, D., Thonnat, M., Bremond, F., & Corvee, E. (2014b). Online parameter tuning for object tracking algorithms. *Image and Vision Computing (IMAVIS)*, *32*, 287–302.

Cristani, M., Raghavendra, R., Del Bue, A., & Murino, V. (2012). Human behavior analysis in video surveillance: a social signal processing perspective. *Neurocomputing*, *100*, 86–97.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 886–893). volume 2.

Easterfield, T. E. (1946). A combinatorial algorithm. *Journal London Mathematical Society*, *21*, 219–226.

Fisher, R. B. (2004). The PETS04 surveillance ground-truth datasets. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)* (pp. 1–5).

Foresti, G., Micheloni, C., Snidaro, L., Remagnino, P., & Ellis, T. (2005). Active video-based surveillance system. *IEEE Signal Processing Magazine (SPM)*, *22*, 25–37.

Gade, R., & Moeslund, T. B. (2014). Thermal cameras and applications: A survey. *Machine Vision and Applications (MVA)*, *25*, 245–262.

Grimson, L., & Stauffer, C. (1999). Adaptive background mixture models for real-time tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2246–2252).

Guan, Y., & Huang, Y. (2015). Multi-pose human head detection and tracking boosted by efficient human head validation using ellipse detection. *Engineering Applications of Artificial Intelligence (EAAI)*, *37*, 181–193.

Hu, W., Tan, T., Wang, L., & Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man and Cybernetics Part C*, *34*, 334–352.

Izadi, M., & Saeedi, P. (2008). Robust region-based background subtraction and shadow removing using color and gradient information. In *International Conference on Pattern Recognition (ICPR)* (pp. 1–5).

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, *82*, 35–45.

Kayumbi, G., Mazzeo, P., Spagnolo, P., Taj, M., & Cavallaro, A. (2008). Distributed visual sensing for virtual top-view trajectory generation in football videos. In *International Conference on Content-based Image and Video Retrieval (CIVR)* (pp. 535–542).

Kuo, C., Huang, C., & Nevatia, R. (2010). Multi-target tracking by on-line learned discriminative appearance models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 685–692).

Kuo, C., & Nevatia, R. (2011). How does person identity recognition help multi-person tracking? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1217–1224).

Li, L., Huang, W., Gu, I. Y. H., Luo, R., & Tian, Q. (2008). An efficient sequential approach to tracking multiple objects through crowds for real-time intelligent CCTV systems. *IEEE Transactions on Systems, Man and Cybernetics Part B*, *38*, 1254–1269.

Li, Y., Huang, C., & Nevatia, R. (2009). Learning to associate: HybridBoosted multi-target tracker for crowded scene. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2953–2960).

Lim, M. K., Tang, S., & Chan, C. S. (2014). iSurveillance: Intelligent framework for multiple events detection in surveillance videos. *Expert Systems With Applications (ESWA)*, *41*, 4704–4715.

Maddalena, L., & Petrosino, A. (2008). A self-organizing approach to background subtraction for visual surveillance applications. *IEEE Transactions on Image Processing (TIP)*, *17*, 1168–1177.

Micheloni, C., Foresti, G., & Snidaro, L. (2005). A network of cooperative cameras for visual-surveillance. *IEE Visual, Image and Signal Processing*, *152*, 205–212.

Moghaddam, B., & Yang, M. (2000). Gender classification with Support Vector Machines. In *IEEE International Conference on Face and Gesture Recognition (FG)* (pp. 306–311).

Novak, C. L., & Shafer, S. A. (1992). Anatomy of a color histogram. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 599–605).

Odobez, J. M., & Yao, J. (2007). Multi-layer background subtraction based on color and texture. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8).

Ojala, T., Pietikainen, M., & Harwood, D. (1994). Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *International Conference on Pattern Recognition (ICPR)* (pp. 582–585). volume 1.

Pavlidis, I., Morellas, V., Tsiamyrtzis, P., & Harp, S. (2001). Urban surveillance systems: From the laboratory to the commercial world. *Proceedings of the IEEE*, *89*, 1478–1497.

Räty, T. D. (2010). Survey on contemporary remote surveillance systems for public safety. *IEEE Transactions on Systems, Man and Cybernetics Part C*, *40*, 493–515.

Reddy, V., Sanderson, C., & Lovell, B. C. (2011). A low-complexity algorithm for static background estimation from cluttered image sequences in surveillance contexts. *EURASIP Journal on Image and Video Processing (JIVP)*, .

Shah, M., Javed, O., & Shafique, K. (2007). Automated visual surveillance in realistic scenarios. *IEEE Multimedia*, *14*, 30–39.

Sobral, A. (2013). BGSLibrary: An OpenCV C++ Background Subtraction Library. In *IX Workshop de Visao Computacional (WVC)*.

Stefano, L. D., Regazzoni, C. S., & Schonfeld, D. (2011). Advanced video-based surveillance. *EURASIP Journal on Image and Video Processing (JIVP)*, .

Szpak, Z. L., & Tapamo, J. R. (2011). Maritime surveillance: Tracking ships inside a dynamic background using a fast level-set. *Expert Systems With Applications (ESWA)*, *38*, 6669–6680.

Vapnik, V., & Cortes, C. (1995). Support-Vector Networks. *Machine Learning*, *20*, 273–297.

Verdant, A., Villard, P., Dupret, A., & Mathias, H. (2011). Three Novell analog-domain algorithms for motion detection in video surveillance. *EURASIP Journal on Image and Video Processing (JIVP)*, .

Walia, G., & Kapoor, R. (2014). Intelligent video target tracking using an evolutionary particle filter based upon improved cuckoo search. *Expert Systems With Applications (ESWA)*, *41*, 6315–6326.

Wang, X. (2013). Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters (PRL)*, *34*, 3–19.

Wong, W. K., Tan, P. N., Loo, C. K., & Lim, W. S. (2009). An effective surveillance system using thermal camera. In *International Conference on Signal Acquisition and Processing (ICSAP)* (pp. 13–17).

Wu, B., & Nevatia, R. (2006). Tracking of multiple, partially occluded humans based on static body part detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 951–958).

Xing, J., Ai, H., & Lao, S. (2009). Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1200–1207).

Xu, L. Q. (2007). Issues in video analytics and surveillance systems: Research / prototyping vs. applications / user requirements. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)* (pp. 10–14).

Zhang, H., & Xu, D. (2006). Fusing color and texture features for background model. In *International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)* (pp. 887–893). volume 4223.

Zhang, S., Wang, J., Wang, Z., Gong, Y., & Liu, Y. (2015). Multi-target tracking by learning local-to-global trajectory models. *Pattern Recognition (PR)*, *48*, 580–590.

Zhao, T., & Nevatia, R. (2004). Tracking multiple humans in crowded environment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 406–413). volume 2.

Zhao, X., Gong, D., & Medioni, G. (2012). Tracking using motion patterns for very crowded scenes. In *European Conference on Computer Vision (ECCV)* (pp. 315–328). volume 2.

Zivkovic, Z. (2004). Improved adaptive gaussian mixture model for background subtraction. In *International Conference on Pattern Recognition (ICPR)* (pp. 28–31).

- Tracking-by-detection based on segmentation, Kalman predictions and LSAP association.
- Occlusion management: SVM kernel metric for GCH + LBP + HOG image features.
- Overall performance near to 85% while tracking under occlusions in CAVIAR dataset.
- Human behavior analysis (exits, loitering, etc.) in naturalistic scenes in shops.
- Real-time multi-camera performance with a processing capacity near to 50fps/camera.