Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 343057, 24 pages
doi:10.1155/2010/343057

Review Article

Background Subtraction for Automated Multisensor Surveillance: A Comprehensive Review

Marco Cristani,[1,2] Michela Farenzena,[1] Domenico Bloisi,[1] and Vittorio Murino[1,2]

[1] *Dipartimento di Informatica, University of Verona, Strada le Grazie 15, 37134 Verona, Italy*
[2] *IIT Istituto Italiano di Tecnologia, Via Morego 30, 16163 Genova, Italy*

Correspondence should be addressed to Marco Cristani, marco.cristani@univr.it

Received 10 December 2009; Accepted 6 July 2010

Academic Editor: Yingzi Du

Background subtraction is a widely used operation in the video surveillance, aimed at separating the expected scene (the background) from the unexpected entities (the foreground). There are several problems related to this task, mainly due to the blurred boundaries between background and foreground definitions. Therefore, background subtraction is an open issue worth to be addressed under different points of view. In this paper, we propose a comprehensive review of the background subtraction methods, that considers also channels other than the sole visible optical one (such as the audio and the infrared channels). In addition to the definition of novel kinds of background, the perspectives that these approaches open up are very appealing: in particular, the multisensor direction seems to be well-suited to solve or simplify several hoary background subtraction problems. All the reviewed methods are organized in a novel taxonomy that encapsulates all the brand-new approaches in a seamless way.

1. Introduction

Video background subtraction represents one of the basic, low-level operations in the video surveillance typical workflow (see Figure 1). Its aim is to operate on the raw video sequences, separating the expected part of the scene (the background, BG), frequently corresponding to the static bit, from the unexpected part (the foreground, FG), often coinciding with the moving objects. Several techniques may subsequently be carried out after the video BG subtraction stage. For instance, tracking may focus only on the FG areas of the scene [1–3]; analogously, target detection and classification may be fastened by constraining the search window only over the FG locations [4]. Further, recognition methods working on shapes (FG silhouettes) are also present in the literature [5, 6]. Finally, the recent coined term of *video analytics* addresses those techniques performing high-level reasoning, such as the detection of abnormal behaviors in a scenery, or the persistent presence of foreground, exploiting low-level operations like the BG subtraction [7, 8].

Video background subtraction is typically an online operation generally composed by two stages, that is, the background initialization, where the model of the background is bootstrapped, and background maintenance (or updating), where the parameters regulating the background have to be updated by online strategies.

The biggest, general problem afflicting the video BG subtraction is that the distinction between the background (the expected part of the scene) and the foreground (the unexpected part) is blurred and cannot fit into the definition given above. For example, one of the problems in video background subtraction methods is the oscillating background: it occurs when elements forming in principle the background, like tree branches in Figure 2, are oscillating. This contravenes the most typical characteristic of the background, that is, that of being static, and bring such items to being labelled as FG instances.

The BG subtraction literature is nowadays huge and multifaceted, with some valid reviews [9–11], and several taxonomies that could be employed, depending on the nature of the experimental settings. More specifically, a first distinction separates the situation in which the sensors (and sensor parameters) are fixed, so that the image view is fixed, and the case where the sensors can move or

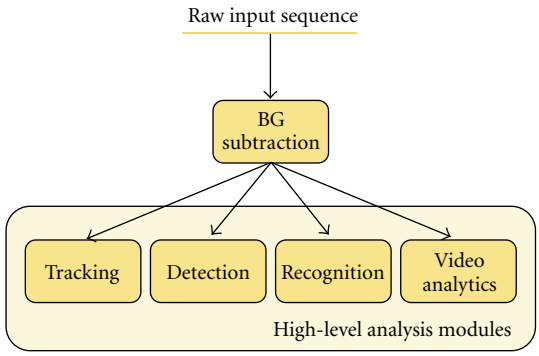


Figure 1: A typical video surveillance workflow: after background subtraction, several, higher-order, analysis procedures may be applied.

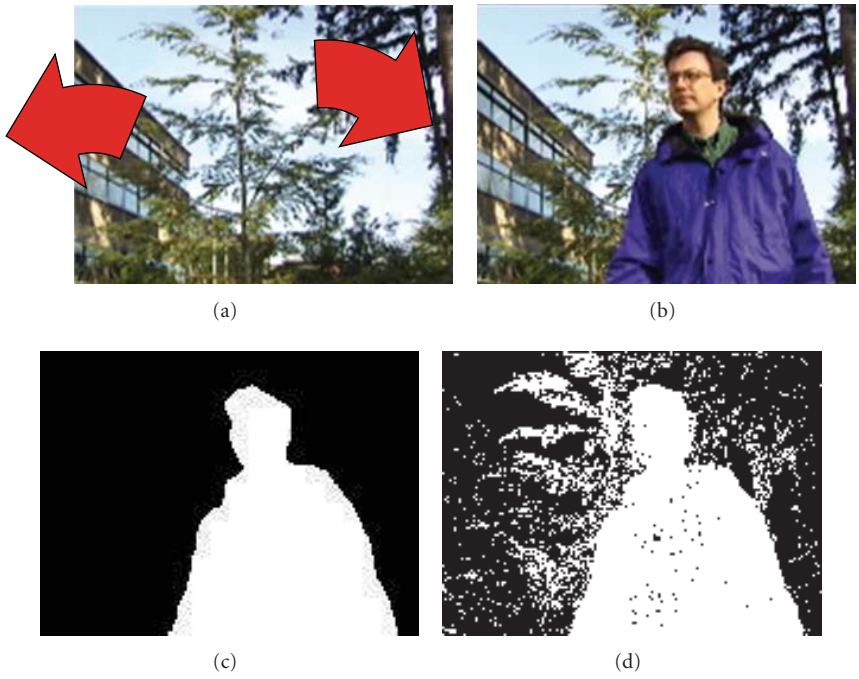


Figure 2: A typical example of ill-posed BG subtraction issue: the oscillating background. (a) A frame representing the background scene, where a tree is oscillating, as highlighted by the arrows. (b) A moving object passes in front of the scene. (c) The ground truth, highlighting only the real foreground object. (d) The result of the background subtraction employing a standard method: the moving branches are detected as foreground.

parameters can change, like cameras mounted on vehicles or PTZ (pan-tilt-zoom) cameras, respectively. In the former case, the scene may be nonperfectly static, especially in the case of an outdoor setting, in which moving foliage or oscillating/repetitively moving entities are present (like flags, water or sea surface): methods in this class try to recover from these noisy sources. In the case of moving sensors, the background is no static any more, and typical strategies aim to individuate the global motion of the scene, separating it from all the other different, local motions that witness the presence of foreground items.

Other taxonomies are more technical, focusing on the algorithmic nature of the approaches, like those separating predictive/nonpredictive [12] or recursive/nonrecursive techniques [13, 14]. In any case, this kind of partitions could not apply to all the techniques present in the literature.

In this paper, we will contribute by proposing a novel, comprehensive, classification of background subtraction techniques, considering not only the mere visual sensor channel, which has been considered by the BG subtraction methods until six years ago. Instead, we will analyze background subtraction *in the large*, focusing on different sensor channels, such as audio and infrared data sources, as well as a combination of multiple sensor channels, like audio + video and infrared + video.

These techniques are very recent and represent the last frontier of the automated surveillance. The adoption of different sensor channels other than video and their careful

association helps in tackling classical unsolved problems for background subtraction.

Considering our multisensor scenario, we thus rewrite the definition of background as *whatever in the scene that is, persistent, under one or more sensor channels*. From this follows the definition of foreground—something that is, not persistent under one ore more sensor channels—and of (multisensor) background subtraction, from here on just background subtraction, unless otherwise specified.

The remainder of the paper is organized as follows. First, we present what are the typical problems that affect the BG subtraction (Section 2) and, afterwards, our taxonomy is described (see Figure 3), using the following structure.

In Section 3, we analyze the BG methods that operate on the sole visible optical (standard video) sensor channel, individuating groups of methods that employ a *single monocular camera*, and approaches where *multiple cameras* are utilized.

Regarding a *single* video stream, *per-pixel* and *per-region* approaches can further be singled out. The rationale under this organization lies in the basic logic entity analyzed by the different methods: in the per-pixel techniques, temporal pixels' profiles are modeled as independent entities. Per-region strategies exploit local analysis on pixel patches, in order to take into account higher-order local information, like edges for instance, also to strengthen the per-pixel analysis. *Per-frame* approaches are based on a reasoning procedure over the entire frame, and are mostly used as support of the other two policies. These classes of approaches can come as integrated multilayer solutions where the FG/BG estimation, made at lower per-pixel level, is refined by the per-region/frame level.

When considering *multiple*, still video, *sensors* (Section 4), we can distinguish between the approaches using sensors in the form of a combined device (such as a stereo camera, where the displacement of the sensors is fixed, and typically embedded in a single hardware platform), and those in which a network of separate cameras, characterized in general by overlapping view fields, is considered.

In Section 5, the approaches devoted to model *audio background* are investigated. Employing audio signals opens up innovative scenarios, where cheap sensors are able to categorize different kind of background situations, highlighting unexpected audio events. Furthermore, in Section 6 techniques exploiting *infrared signals* are considered. They are particularly suited when the illumination of the scene is very scarce. This concludes the approaches relying on a single sensor channel.

The subsequent part analyzes how the single sensor channels, possibly modeled with more than one sensor, could be jointly employed through fusion policies in order to estimate *multisensor background models*. They inherit the strengths of the different sensor channels, and minimize the drawbacks typical of the single separate channels. In particular, we will investigate in Section 7 the approaches that fuse infrared + video and audio + video signals (see Figure 3).

This part concludes the proposed taxonomy and is followed by the summarizing Section 8, where the typical problems of the BG subtraction are discussed, individuating the reviewed approaches that cope with some of them. Then, for each problem, we will give a sort of recipe, distilled from all of the approaches analyzed, that indicates how that specific problem can be solved. These considerations are summed up in Table 1.

Finally, a conclusive part, (Section 9), closes the survey, envisaging which are the unsolved problems, and discussing what are the potentialities that could be exploited in the future research.

As a conclusive consideration, it is worth noting that our paper will not consider solely papers that focus in their entirety on a BG subtraction technique. Instead, we decide to include those works where the BG subtraction represents a module of a structured architecture and that bring advancements in the BG subtraction literature.

2. Background Subtraction's Key Issues

Background subtraction is a hard task as it has to deal with different and variable issues, depending on the kind of environment considered. In this section, we will analyze such issues following the idea adopted for the development of the "Wallflower" dataset (http://research.microsoft.com/en-us/um/people/jckrumm/WallFlower/TestImages.htm) presented in [15]. The dataset consists of different video sequences that is, olate and portray single issues that make the BG/FG discrimination difficult. Each sequence contains a frame which serves as test, and that is, given together with the associated ground truth. The ground truth is represented by a binary FG mask, where 1 (white) stands for FG. It is worth noting that the presence of a test frame indicates that *in that frame* a BG subtraction issue occurs; therefore, the rest of the sequence cannot be strictly considered as an instance of a BG subtraction problem.

Here, we reconsider these same sequences together with new ones showing problems that are not taken into account in the Wallflower work. Some sequences portray also problems which rarely have been faced in the BG subtraction literature. In this way, a very comprehensive list of BG subtraction issues is given, associated with representative sequences (developed by us or already publicly available) that can be exploited for testing the effectiveness of novel approaches.

For the sake of clarity, from now on we assume as false positive a FG entity which is identified as BG, and viceversa.

Here is the list of problems and their relative representative sequences (http://profs.sci.univr.it/~cristanm/BGsubtraction/videos) (see Figure 4):

Moved Object [15]. A background object can be moved. Such object should not be considered part of the foreground forever after, so the background model has to adapt and understand that the scene layout may be physically updated. This problem is tightly connected with that of the sleeping person (see below), where a FG object stand still in the scene and, erroneously, becomes part of the scene. The sequence portrays a chair that is, moved in a indoor scenario.

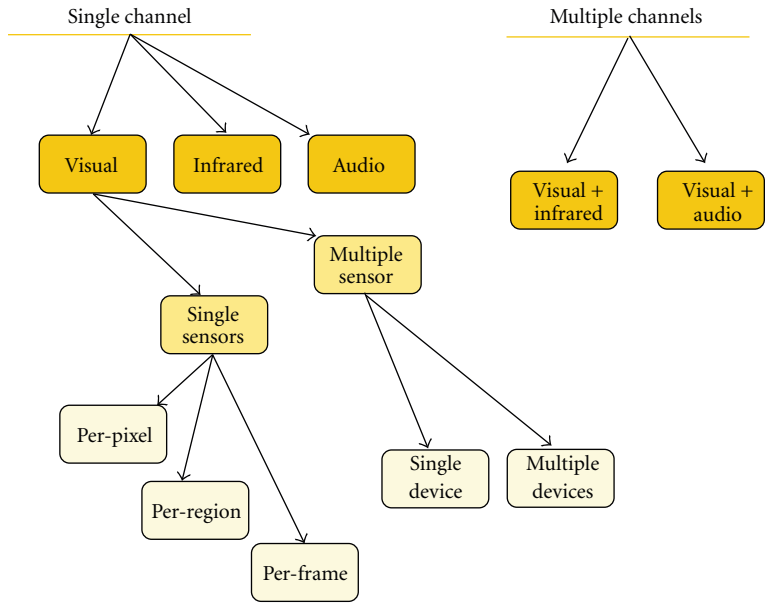


Figure 3: Taxonomy of the proposed background subtraction methods.

Time of Day [15]. Gradual illumination changes alter the appearance of the background. In the sequence the evolution of the illumination provokes a global appearance change of the BG.

Light Switch [15]. Sudden changes in illumination alter the appearance of the background. This problem is more difficult than the previous one, because the background does evolve with a characteristic that is, typical of a foreground entity, that is, being unexpected. In their paper [15], the authors present a sequence where a global change in the illumination of a room occurs. Here, we articulate this situation adding the condition where the illumination change may be local. This situation may happen when street lamps are turned on in an outdoor scenario; another situation may be that of an indoor scenario, where the illumination locally changes, due to different light sources. We name such problem, and the associated sequence, *Local light switch*. The sequence shows an indoor scenario, where a dark corridor is portrayed. A person moves between two rooms, opening and closing the related doors. The light in the rooms is on, so the illumination spreads out over the corridor, locally changing the visual layout. A background subtraction algorithm has to focus on the moving entity.

Waving Trees [15]. Background can vacillate, globally and locally, so the background is not perfectly static. This implies that the movement of the background may generate false positives (movement is a property associated to the FG). The sequence, depicted also in Figure 2, shows a tree that is, moved continuously, simulating an oscillation in an outdoor situation. At some point, a person comes. The algorithm has to highlight only the person, not the tree.

Camouflage [15]. A pixel characteristic of a foreground object may be subsumed by the modeled background, producing a false negative. The sequence shows a flickering monitor that alternates shades of blue and some white regions. At some point, a person wearing a blue shirt moves in front of the monitor, hiding it. The shirt and the monitor have similar color information, so the FG silhouette tends do be erroneously considered as a BG entity.

Bootstrapping [15]. A training period without foreground objects is not always available in some environments, and this makes bootstrapping the background model hard. The sequence shows a coffee room where people walk and stay standing for a coffee. The scene is never empty of people.

Foreground Aperture [15]. When a homogeneously colored object moves, changes in the interior pixels cannot be detected. Thus, the entire object may not appear as foreground, causing false negatives. In the Wallflower sequence, this situation is made even extreme. A person is asleep at his desk, viewed from the back. He wakes up and *slowly* begins to move. His shirt is uniformly colored.

Sleeping Foreground. A foreground object that becomes motionless has to be distinguished from the background. In [15], this problem has not been considered because it implies the knowledge of the foreground. Anyway, this problem is similar to that of the "moved object". Here, the difference is that the object that becomes still does not belong to the scene. Therefore, the reasoning for dealing with this problem may be similar to that of the "moved object". Moreover, this problem occurs very often in the surveillance situations, as witnessed by our test sequence. This sequence portrays a crossing road with traffic lights, where the cars move and stop. In such a case, the cars have not to be marked as background.

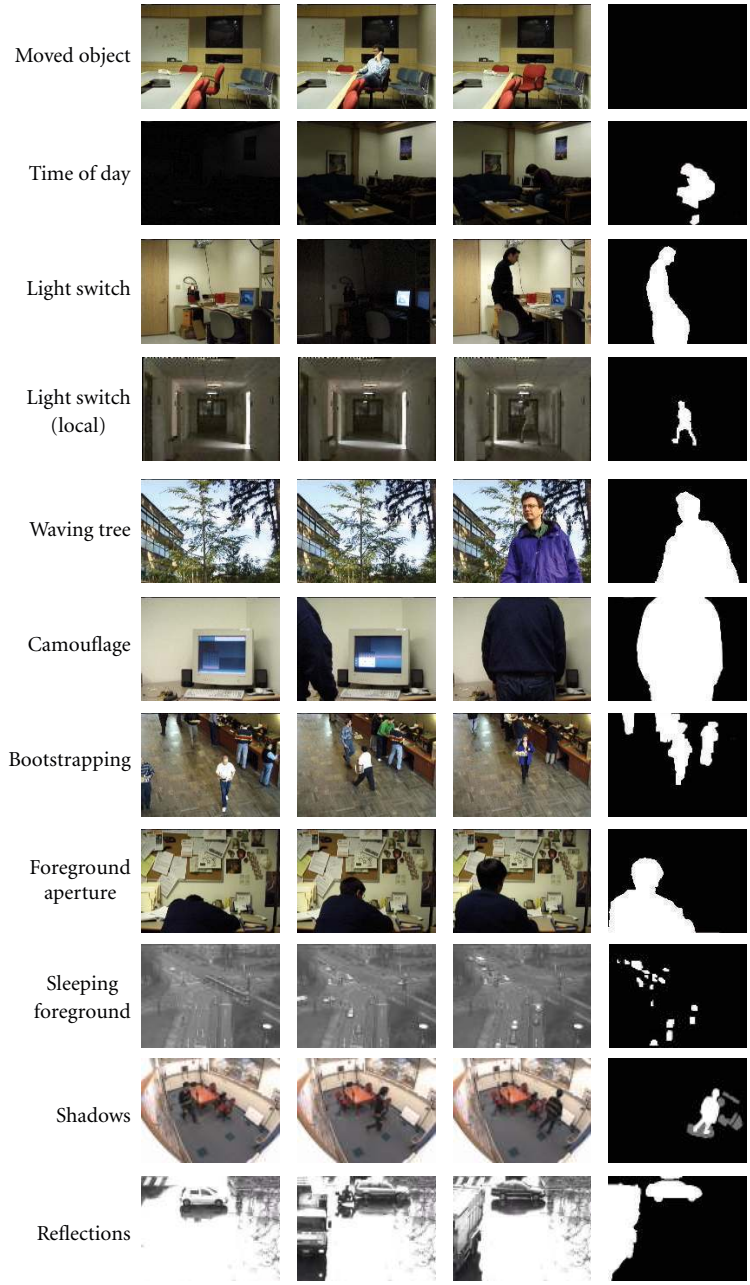


Figure 4: Key problems for the BG subtraction algorithms. Each situation corresponds to a row in the figure, the images in the first two column (starting from left) represent two frames of the sequence, the images in the third column represent the test image, and the images in the fourth column represent the ground truth.

Shadows. Foreground objects often cast shadows that appear different from the modeled background. Shadows are simply erratic and local changes in the illumination of the scene, so they have not to be considered FG entities. Here we consider a sequence coming from the ATON project (http://cvrr.ucsd.edu/aton/testbed/), depicting an indoor scenario, where a person moves, casting shadows on the floor and on the walls. The ground truth presents two labels: one for the foreground and one for the shadows.

Reflections. the scene may reflects foreground instances, due to wet or reflecting surfaces, such as the floor, the road, windows, glasses, and so for, and such entities have not to be classified as foreground. In the literature, this problem has been never explicitly studied, and it has been usually aggregated with that of the shadows. Anyway, reflections are different from shadows, because they retain edge information that is, absent in the shadows. We present here a sequence where a traffic road intersection is monitored.

Table 1: A summary of the methods discussed in this paper, associated with the problems they solve. The meaning of the abbreviations is reported in the text.

	MO	TD	LS	LLS	WT	C	B	FGA	SFG	SH	R
Per-pixel		√			√		√			√	
Per-region		√			√	√	√	√		√	
Per-frame		√	√				√				
Multistage		√	√	√		√	√			√	
Multicamera		√	√	√		√		√		√	
Infrared-sensor										√	
Infrared + video										√	
Infrared + video									√		

The floor is wet and the shining sun provokes reflections of the passing cars.

In the following section, we will consider these situations with respect to how the different techniques present in the literature solve them (we explicitly refer to those approaches that consider the presented test sequences) or may help in principle to reach a good solution (in this case, we infer that a good solution is given for a problem when the sequence considered are similar to those of the presented dataset).

Please note that the Wallflower sequences contain only video data, and so all the other new sequences. Therefore, for the approaches that work on other sensor channels, the capability to solve one of these problems will be based on results applied on data sequences that present analogies with the situations portrayed above.

3. Single Monocular Video Sensor

In a single camera setting, background subtraction focuses on a pixel matrix that contains the data acquired by a black/white or color camera. The output is a binary mask which highlights foreground pixels. In practice, the process consists in comparing the current frame with the background model, individuating as foreground pixels those not belonging to it.

Different classifications of BG subtraction methods for monocular sensor settings have been proposed in literature. In [13, 14], the techniques are divided into recursive and nonrecursive ones, where recursive methods maintain a single background model that is, updated using each new coming video frame. Nonrecursive approaches maintain a buffer with a certain quantity of previous video frames and estimate a background model based solely on the statistical properties of these frames.

A second classification [12] divides existing methods in predictive and nonpredictive. Predictive algorithms model a scene as a time series and develop a dynamic model to evaluate the current input based on the past observations. Nonpredictive techniques neglect the order of the input observations and build a probabilistic representation of the observations at a particular pixel.

However, the above classifications do not cover the entire range of existent approaches (actually, there are techniques that contain predictive and nonpredictive parts), and does not give hints on the capabilities of each approach.

The Wallflower paper [19] inspired us a different taxonomy, similar to the one proposed in [20], that fills this gap. Such work actually proposes a method that works on different spatial levels: per-pixel, per-region, and per-frame. Each level taken alone has its own advantages and is prone to well defined key problems; moreover, each level individuates several approaches in the literature. Therefore, individuating an approach as working solely in a particular level makes us aware of what problems that approach can solve. For example, considering every temporal pixel evolution as an independent process (so addressing the per-pixel level), and ignoring information observed at the other pixels (so without performing any per-region/frame reasoning) cannot be adequate for managing the light switch problem. This partition of the approaches into spatial logic levels of processing (pixel, region, and frame) is consistent with the nowadays BG subtraction state of the art, permitting to classify all the existent approaches.

Following these considerations, our taxonomy organizes the BG subtraction methods into three classes.

(i) *Per-Pixel Processing.* The class of per-pixel approaches is formed by methods that perform BG/FG discrimination by considering each pixel signal as an independent process. This class of approaches is the most adopted nowadays, due to the low computational effort required.

(ii) *Per-Region/Frame Processing.* Region-based algorithms relax the per-pixel independency assumption, thus permitting local spatial reasoning in order to minimize false positive alarms. The underlying motivations are mainly twofold. First, pixels may model parts of the background scene which are locally oscillating or moving slightly, like leafs or flags. Therefore, the information needed to capture these BG phenomena has not to be collected and evaluated over a single pixel location, but on a larger support. Second, considering the neighborhood of a pixel permits to assess useful analysis, such as edge extraction or histogram computation. This provides a more robust description of the visual appearance of the observed scene.

(iii) *Per-Frame Processing.* Per-frame approaches extend the local support of the per-region methods to the entire frame, thus facing global problems like the light switch.

3.1. Per-Pixel Processes. In order to ease the reading, we group together similar approaches, considering the most important characteristics that define them. This permits also to highlight in general pros and cons of multiple approaches.

3.1.1. Early Attempts of BG Subtraction. To the best of our knowledge, the first attempt to implement a background subtraction model for surveillance purposes is the one in [21], where the differencing of adjacent frames in a video sequence are used for object detection in stationary cameras. This simple procedure is clearly not adapt for long-term analysis, and suffers from many practical problems (one for all, it does not highlight the entire FG appearance, due to the overlapping between moving objects across frames).

3.1.2. Monomodal Approaches. Monomodal approaches assumes that the features that characterize the BG values of a pixel location can be segregated in a single compact support. One of the first and widely adopted strategy was proposed in the surveillance system Pfinder [22], where each pixel signal $z^{(t)}$ is modeled in the YUV space by a simple mean value, updated on-line. At each time step, the likelihood of the observed pixel signal, given an estimated mean, is computed and a FG/BG labeling is performed.

A similar approach has been proposed in [23], exploiting a running Gaussian average. The background model is updated if a pixel is marked as foreground for more than m of the last M frames, in order to compensate for sudden illumination changes and the appearance of static new objects. If a pixel changes state from FG to BG frequently, it is labeled as a high-frequencies background element and it is masked out from inclusion in the foreground.

Median filtering sets each color channel of a pixel in the background as modeled by the median value, obtained from a buffer of previous frames. In [24], a recursive filter is used to estimate the median, achieving a high computational efficiency and robustness to noise. However, a notable limit is that it does not model the variance associated to a BG value.

Instead of independently estimating the median of each channel, the medoid of a pixel can be estimated from the buffer of video frames as proposed in [25]. The idea is to consider color channels together, instead of treating each color channel independently. This has the advantage of capturing the statistical dependencies between color channels.

In W^4 [26, 27], a pixel is marked as foreground if its value satisfies a set of inequalities, that is

$$\left|M - z^{(t)}\right| > D \vee \left|N - z^{(t)}\right| > D, \tag{1}$$

where the (per-pixel) parameters M, N, and D represent the minimum, maximum, and largest interframe absolute difference observable in the background scene, respectively. These parameters are initially estimated from the first few seconds of a video and are periodically updated for those parts of the scene not containing foreground objects.

The drawback of these models are that only monomodal background are taken into account, thus ignoring all the situations where multimodality in the BG is present. For example, considering a water surface, each pixel has at least a bimodal distribution of colors, highlighting the sea and the sun reflections.

3.1.3. Multimodal Approaches. One of the first approaches dealing with multimodality is proposed in [28], where a mixture of Gaussians is incrementally learned for each pixel. The application scenario is the monitoring of an highway, and a set of heuristics for labeling the pixels representing the road, the shadows and the cars are proposed.

An important approach that introduces a parametric modeling for multimodal background is the Mixture of Gaussians (MoG) model [29]. In this approach, the pixel evolution is statistically modeled as a multimodal signal, described using a time-adaptive mixture of Gaussian components, widely employed in the surveillance community. Each Gaussian component of a mixture describes a gray level interval observed at a given pixel location. A weight is associated to each component, mirroring the confidence of portraying a BG entity. In practice, the higher the weight, the stronger the confidence, and the longer the time such gray level has been recently observed at that pixel location. Due to the relevance assumed in the literature and the numerous proposed improvements, we perform here a detailed analysis of this approach.

More formally, the probability of observing the pixel value $z^{(t)}$ at time t is

$$P\left(z^{(t)}\right) = \sum_{r=1}^{R} w_r^{(t)} \mathcal{N}\left(z^{(t)} \mid \mu_r^{(t)}, \sigma_r^{(t)}\right), \tag{2}$$

where $w_r^{(t)}$, $\mu_r^{(t)}$ and $\sigma_r^{(t)}$ are the mixing coefficients, the mean, and the standard deviation, respectively, of the rth Gaussian $\mathcal{N}(\cdot)$ of the mixture associated with the signal at time t. The Gaussian components are ranked in descending order using the w/σ value: the most ranked components represent the "expected" signal, or the background.

At each time instant, the Gaussian components are evaluated in descending order to find the first matching with the observation acquired (a *match* occurs if the value falls within 2.5σ of the mean of the component). If no match occurs, the least ranked component is discarded and replaced with a new Gaussian with the mean equal to the current value, a high variance σ_{init}, and a low mixing coefficient w_{init}. If r_{hit} is the matched Gaussian component, the value $z^{(t)}$ is labeled FG if

$$\sum_{r=1}^{r_{\text{hit}-1}} w_r^{(t)} > T, \tag{3}$$

where T is a standard threshold. The equation that drives the evolution of the mixture's weight parameters is the following:

$$w_r^{(t)} = (1-\alpha) w_r^{(t-1)} + \alpha M^{(t)}, \quad 1 \le r \le R, \tag{4}$$

(a) (b)

Figure 5: A near infrared image (a) from CBSR dataset [16, 17] and a thermal image (b) from Terravic Research Infrared Database [17, 18].

where $M^{(t)}$ is 1 for the matched Gaussian (indexed by r_{hit}) and 0 for the others, and α is the learning rate. The other parameters are updated as follows:

$$\begin{aligned}\mu_{r_{\text{hit}}}^{(t)} &= (1-\rho)\mu_{r_{\text{hit}}}^{(t-1)} + \rho z^{(t)},\\ \sigma_{r_{\text{hit}}}^{2\,(t)} &= (1-\rho)\sigma_{r_{\text{hit}}}^{2\,(t-1)} + \rho\left(z^{(t)} - \mu_{r_{\text{hit}}}^{(t)}\right)^T\left(z^{(t)} - \mu_{r_{\text{hit}}}^{(t)}\right),\end{aligned} \tag{5}$$

where $\rho = \alpha\mathcal{N}(z^{(t)} \mid \mu_{r_{\text{hit}}}^{(t)}, \sigma_{r_{\text{hit}}}^{(t)})$. It is worth noting that the higher the adaptive rate α, the faster the model is "adapted" to signal changes. In other words, for a low learning rate, MoG produces a wide model that has difficulty in detecting a sudden change to the background (so, it is prone to the light switch problem, global and local). If the model adapts too quickly, slowly moving foreground pixels will be absorbed into the background model, resulting in a high false negative rate (the problem of the foreground aperture).

MoG has been further improved by several authors, see [30, 31]. In [30], the authors specify (i) how to cope with color signals (the original version was proposed for gray values), proposing a normalization of the RGB space taken from [12], (ii) how to avoid overfitting and underfitting (values of the variances too low or too high), proposing a thresholding operation, and (iii) how to deal with sudden and global changes of the illumination, by changing the learning rate parameter. For the latter, the idea is that if the foreground changes from one frame to another more than the 70%, the learning rate value grows up, in order to permit a faster evolution of the BG model. Note that this improvement adds global (per-frame) reasoning to MoG, so it does not belong properly to the class of per-pixel approaches.

In [31], the number of Gaussian components is automatically chosen, using a Maximum A-Posteriori (MAP) test and employing a negative Dirichlet prior.

Even if per-pixel algorithms are widely used for their excellent compromise between accuracy and speed (in computational terms), these techniques present some drawbacks, mainly due to the interpixel independency assumption. Therefore, any situation that needs a global view of the scene in order to perform a correct BG labeling is lost, usually causing false positives. Examples of such situations are sudden changes in the chromatic aspect of the scene, due to the weather evolution or local light switching.

3.1.4. Nonparametric Approaches. In [32], a nonparametric technique estimating the per-pixel probability density function using the kernel density estimation (KDE) [33] technique is developed (KDE method is an example of Parzen window estimate, [34]). This faces the situation where the pixel values" density function is complex and cannot be modeled parametrically, so a non-parametric approach able to handle arbitrary densities is more suitable. The main idea is that an approximation of the background density can be given by the histogram of the most recent values classified as background values. However, as the number of samples is necessarily limited, such an approximation suffers from significant drawbacks: the histogram might provide poor modeling of the true pdf, especially for rough bin quantizations, with the tails of the true pdf often missing. Actually, KDE guarantees a smoothed and continuous version of the histogram. In practice, the background pdf is given as a sum of Gaussian kernels centered in the most recent n background values, b_i

$$P\left(z^{(t)}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(z^{(t)} - b_i, \Sigma_t\right). \tag{6}$$

In this case, each Gaussian describes one sample data, and not a whole mode as in [29], with n in the order of 100, and covariance fixed for all the samples and all the kernels. The classification of $z^{(t)}$ as foreground is assumed when $P(z^{(t)}) < T$. The parameters of the mixtures are updated by changing the buffer of the background values in FIFO order by selective update, and the covariance (in this case, a diagonal matrix) is estimated in the time domain by analyzing the set of differences between two consecutive values. In [32], such model is duplicated: one model is employed for a long-term background evolution modeling (for example dealing with the illumination evolution in a outdoor scenario) and the other for the short-term modeling

(for flickering surfaces of the background). Intersecting the estimations of the two models gives the first stage results of detection. The second stage of detection aims at suppressing the false detections due to small and unmodelled movements of the scene background that cannot be observed employing a per-pixel modeling procedure alone. If some parts of the background (a tree branch, for example) moves to occupy a new pixel, but it is not part of the model for that pixel, it will be detected as a foreground object. However, this object will have a high probability to be a part of the background distribution at its original pixel location. Assuming that only a small displacement can occur between consecutive frames, a detected FG pixel is evaluated as caused by a background object that has moved by considering the background distributions in a small neighborhood of the detection area. Considering this step, this approach could also be intended as per-region.

In their approach, the authors also propose a method for dealing with the shadows problem. The idea is to separate the color information from the lightness information. Chromaticity coordinates [35] help in suppressing shadows, but loses lightness information, where the lightness is related to the difference in whiteness, blackness and grayness between different objects. Therefore, the adopted solution considers $S = R + G + B$ as a measure of lightness, where R, G and B are the intensity values for each color channel of a given pixel. Imposing a range on the ratio between a BG pixel value and its version affected by a shadow permits to perform a good shadow discrimination. Please note that, in this case, the shadow detection relies on a pure per-pixel reasoning.

Concerning the computational efforts of the per-pixel processes, in [9] a good analysis is given: speed and memory usage of some widely used algorithms are taken into account. Essentially, monomodal approaches are generally the fastest, while multimodal and non-parametric techniques exhibit higher complexity. Regarding the memory usage, non-parametric approaches are the most demanding, because they need to collect for each pixel a statistics on the past values.

3.2. Per-Region Processes. Region-level analysis considers a higher level representation, modeling also interpixel relationships, allowing a possible refinement of the modeling obtained at the pixel level. Region-based algorithms usually consider a local patch around each pixel, where local operations may be carried out.

3.2.1. Nonparametric Approaches. This class could include also the approach of [32], above classified as per-pixel, since it incorporats a part of the technique (the false suppression step) that is, inherently per-region.

A more advanced approach using adaptive kernel density estimation is proposed in [12]. Here, the model is genuinely region-based: the set of pixels values needed to compute the histogram (i.e., the nonparametric density estimate for a pixel location) is collected over a local spatial region around that location, and not exclusively on the past values of that pixel.

3.2.2. Texture- and Edge-Based Approaches. These approaches exploit the spatial local information for extracting structural information such as edges or textures. In [36], video sequences are analyzed by dividing the scene in overlapped squared patches. Then, intensity and gradient kernel histograms are built for each patch. Roughly speaking, intensity (gradient) kernel histograms count pixel (edge) values as weighted entities, where the weight is given by a Gaussian kernel response. The Gaussian kernel, applied on each patch, gives more importance to the pixel located in the center. This formulation gives invariance to illumination changes and shadows because the edge information helps in discriminating a FG occluding object, that introduces different edge information in the scene, and a (light) shadow, that only weakens the BG edge information.

In [37], a region model describing local texture characteristics is presented through a modification of the Local Binary Patterns [38]. This method considers for each pixel a fixed circular region and calculates a binary pattern of length N where each ordered value of the pattern is 1 if the difference between the center and a particular pixel lying on the circle is larger than a threshold. This pattern is calculated for each neighboring pixel that lies in the circular region. Therefore, a histogram of binary patterns is calculated. This is done for each frame and, subsequently, a similarity function among histograms is evaluated for each pixel, where the current observed histogram is compared with a set of K weighted existing models. Low-weighted models stand for FG, and vice versa. The model most similar to the histogram observed is the one that models the current observation, so increasing its weight. If no model explains the observation, the pixel is labeled as FG, and a novel model is substituted with the least supported one. The mechanism is similar to the one used for per-pixels BG modeling proposed in [29].

The texture analysis for BG subtraction is considered also in [39], where it is proposed a combined pixel-region model where the color information associated to a pixel is defined in a photometric invariant space, and the structural region information derives from a local binary pattern descriptor, defined in the pixel's neighborhood area. The two aspects are linearly combined in a whole signature that lives in a multimodal space, which is modeled and evaluated similarly to MoG. This model results particularly robust to shadows.

Another very similar approach is presented in [40], where color and gradient information are explicitly modeled as time adaptive Gaussian mixtures.

3.2.3. Sampling Approaches. The sampling approaches evaluate a wide local area around each pixel to perform complex analysis. Therefore, the information regarding the spatial support is collected through sampling, which in some cases permits to fasten the analysis.

In [41], the pixel-region mixing is carried out with a spatial sampling mechanism, that aims at producing a finer BG model by propagating BG pixels values in a local area. This principle resembles a region growing segmentation algorithm, where the statistics of an image region is built by considering all the belonging pixels. In this way, regions affected by a local, small chromatic variation (due to a cloudy

weather or shadows, for example), become less sensitive to the false positives. The propagation of BG samples is done with a particle filter policy, and a pixel values with higher likelihood of being BG is propagated farer in the space. As per-pixel model, a MoG model is chosen. The drawback of the method is that it is computational expensive, due to the particle filtering sampling process.

In [42] a similar idea of sampling the spatial neighborhood for refining the per-pixel estimate is adopted. The difference here lies in the per-pixel model, that is, non-parametric, and it is based on a Parzen windows-like process. The model updating relies on a random process that substitutes old pixel values with new ones. The model has been compared favorably with the MoG model of [31] with a small experimental dataset.

3.2.4. BG Subtraction Using a Moving Camera. The approaches dealing with moving cameras focus mainly on compensating the camera ego-motion, checking if the statistics of a pixel can be matched with the one present in a reasonable neighborhood. This occurs through the use of homographies or 2D affine transformations of layered representations of the scene.

Several methods [43–46] well apply to scenes where the camera center does not translate, that is, when using of PTZ cameras (pan, tilt, or zoom motions). Another favorable scenario is when the background can be modeled by a plane. When the camera may translate and rotate, other strategies have been adopted.

In the plane + parallax framework [47–49], a homography is first estimated between successive image frames. The registration process removes the effects of camera rotation, zoom, and calibration. The residual pixels correspond either to moving objects or to static 3D structures with large depth variance (parallax pixels). To estimate the homographies, these approaches assume the presence of a dominant plane in the scene, and have been successfully used for object detection in aerial imagery where this assumption is usually valid.

Layer-based methods [50, 51] model the scene as piecewise planar scenes, and cluster segments based on some measure of motion coherency.

In [52], a layer-based approach is explicitly suited for background subtraction from moving cameras but report low performance for scenes containing significant parallax (3D scenes).

Motion segmentation approaches like [53, 54] sparsely segment point trajectories based on the geometric coherency of the motion.

In [55], a technique based on sparse reasoning is presented, which also deals with rigid and nonrigid FG objects of various size, merged in a full 3D BG. The underlying assumptions regard the use of an orthographic camera model and that the background is the spatially dominant rigid entity in the image. Hence, the idea is that the trajectories followed by sparse points of the BG scene lie in a three-dimensional subspace, estimated through RANSAC, so allowing to highlight outlier trajectories as FG entities, and to produce a sparse pixel FG/BG labeling. Per-pixel labels are then coupled together through the use of a Markov Random Field (MRF) spatial prior. Limitations of the model concern the considered approximation of the camera model, affine instead of fully perspective, but, experimentally, it has been shown not to be very limiting.

3.2.5. Hybrid Foreground/Background Models for BG Subtraction. These models includes in the BG modeling a sort of knowledge of the FG, so they may not be classified as pure BG subtraction methods. In [20], a BG model competes with an explicit FG model in providing the best description of the visual appearance of a scene. The method is based on a maximum a posteriori framework, which exhibits the product of a likelihood term and a prior term, in order to classify a pixel as FG or BG. The likelihood term is obtained exploiting a ratio between nonparametric density estimations describing the FG and the BG, respectively, and the prior is given by employing an MRF that models spatial similarity and smoothness among pixels. Note that, other than the MRF prior, also the non-parametric density estimation (obtained using the Parzen Windows method) works on a region level, looking for a particular signal intensity of the pixel in an isotropic region defined on a joint spatial and color domain.

The idea of considering a FG model together with a BG model for the BG subtraction has been also taken into account in [56], where a pool of local BG features is selected at each time step in order to maximize the discrimination from the FG objects. A similar approach has been taken into account in [57], where the authors propose a boosting approach which selects the best features for separating BG and FG.

Concerning the computational efforts, per-region approaches exhibit higher complexity, both in space and in time, than the per-pixel ones. Anyway, the most papers claim real-time performances.

3.3. Per-Frame Approaches. These approaches extend the local area of refinement of the per-pixel analysis to being the entire frame. In [58], a graphical model is used to adequately model illumination changes of a scene. Even if results are promising, it is worth noting that the method has not be evaluated in its on-line version, nor it works in real-time; further, illumination changes should be global and pre-classified in a training section.

In [59], a per-pixel BG model was chosen from a set of pre-computed ones in order to minimize massive false alarm.

The method proposed in [60] captures spatial correlations by applying principal component analysis [34] to a set of N_L video frames that do not contain any foreground objects. This results in a set of basis functions, whose the first d are required to capture the primary appearance characteristics of the observed scene. A new frame can then be projected into the eigenspace defined by these d basis functions and then back projected into the original image space. Since the basis functions only model the static part of the scene when no foreground objects are present, the back projected image will not contain any foreground objects. As such, it can be used as a background model.

The major limitation of this approach lies just on the original hypothesis of absence of foreground objects to compute the basis functions which is not always possible. Moreover, it is also unclear how the basis functions can be updated over time if foreground objects are going to be present in the scene.

Concerning the computational efforts, per-frames approaches usually are based on a training step and classification step. The training part is carried out in a offline fashion, while the classification part is well suited for a real-time usage.

3.4. Multistage Approaches. The multistage approaches consist in those techniques that are formed by several serial heterogeneous steps, that thus cannot be included properly in any of the classes seen before.

In Wallflower [15], a 3-stage algorithm that operates respectively at pixel, region and frame level is presented.

At the pixel level, a couple of BG models is maintained for each pixel independently: both the models are based on a 40-coefficients, one-step Wiener filter, where the (past) values taken into account are the predicted values by the filter in one case, and the observed values in the other. A double check against these two models is performed at each time step: the current pixel value is considered as BG if it differs less than 4 times the expected squared prediction error calculated using the two models.

At the region level, a region growing algorithm is applied. It essentially closes the possible holes (false negative) in the FG if the signal values in the false negative locations are similar to the values of the surrounding FG pixels. At the frame level, a set of global BG models is finally generated. When a big portion of the scene is suddenly detected as FG, the best model is selected, that is, the one that minimizes the amount of FG pixels.

A similar, multilevel approach has been presented in [61], where the problem of the local/global light switch is taken into account. The approach lies on a segmentation of the background [62] which segregates portions of the scene where the chromatic aspect is homogeneous and evolves uniformly. When a background region suddenly changes its appearance, it is considered as a BG evolution instead of a FG appearance. The approach works well when the regions in the scene are few and wide. Conversely, the performances are poor when the scene is oversegmented, that in general occurs for outdoor scenes.

In [63], the scene is partitioned using a quadtree structure, formed by minimal average correlation energy (MACE) filters. Starting with large-sized filters (32×32 pixels), 3 levels of smaller filters are employed, until the lower level formed by 4×4 filters. The proposed technique aims at avoiding false positives: when a filter detects the FG presence on more than 50% of its area, the analysis is propagated to the 4 children belonging to the lower level, and in turn to the 4-connected neighborhood of each one of the children. When the analysis reaches the lowest (4×4) level and FG is still discovered, the related set of pixels are marked as FG. Each filter modeling a BG zone is updated, in order to deal with slowly changing BG. The method is slow and no real-time implementation is presented by the authors, due to the computation of the filters' coefficients.

This computational issue has been subsequently solved in [64]. Given the same quadtree structure, instead of entirely analyzing each zone covered by a filter, only one pixel is randomly sampled and analyzed for each region (filter) at the highest level of the hierarchy. If no FG is detected, the analysis stops; otherwise, the analysis is further propagated on the 4 children belonging to the lower level, down to reach the lowest one. Here, in order to get the fine boundaries of the BG silhouette, a 4-connected neighborhood region growing algorithm is performed on each of the FG children. The exploded quadtree is used as default structure for the next frame in order to cope efficiently with the overlap among FG regions between consecutive frames.

In [65], a nonparametric, per pixel FG estimation is followed by a set of morphological operations in order to solve a set of BG subtraction common issues. These operations evaluate the joint behavior of similar and proximal pixel values by connected-component analysis that exploits the chromatic information. In this way, if several pixels are marked as FG, forming a connected area with possible holes inside, the holes can be filled in. If this area is very large, the change is considered as caused by a fast and global BG evolution, and the entire area is marked as BG.

All the multistage approaches require high computational efforts, if compared with the previous analysis paradigms. Anyway, in all the aforementioned papers the multistage approaches are claimed to be functioning in a real-time setting.

3.5. Approaches for the Background Initialization. In the realm of the BG subtraction approach in a monocular video scenario, a quite relevant aspect is the one of the background initialization, that is, how a background model has to be bootstrapped. In general, all of the presented methods discard the solution of computing a simple mean over all the frames, because it produces an image that exhibits blending pixel values in areas of foreground presence. A general analysis regarding the blending rate and how it may be computed is present in [66].

In [67], the background initial values are estimated by calculating the median value of all the pixels in the training sequence, assuming that the background value in every pixel location is visible more than 50% of the time during the training sequence. Even if this method avoids the blending effects of the mean, the output of the median will contains large error when this assumption is false.

Another proposed work [68], called adaptive smoothness method, avoids the problem of finding intervals of stable intensity in the sequence. Then, using some heuristics, the longest stable value for each pixel is selected and used as the value that most likely represents the background.

This method is similar to the recent Local Image Flow algorithm [69], which generates background values' hypotheses by locating intervals of relatively constant intensity, and weighting these hypotheses by using local motion

information. Unlike most of the proposed approaches, this method does not treat each pixel value sequence as an i.i.d. (independent identically distributed) process, but it considers also information generated by the neighboring locations.

In [62], a hidden Markov model clustering approach was proposed in order to consider homogeneous compact regions of the scene whose chromatic aspect does uniformly evolve. The approach fits a HMM for each pixel location, and the clustering operates using a similarity distance which weights more heavily the pixel values portraying BG values.

In [70], an inpainting-based approach for BG initialization is proposed: the idea is to apply a region-growing spatiotemporal segmentation approach, which is able expand a safe, local, BG region by exploiting perceptual similarity principles. The idea has been further improved in [71], where the region growing algorithm has been further developed, adopting graph-based reasoning.

3.6. Capabilities of the Approaches Based on a Single Video Sensor. In this section, we summarize the capabilities of the BG subtraction approaches based on a monocular video camera, by considering their abilities in solving the key problems expressed in Section Problems.

In general, whatever approach which permits an adaptation of the BG model can deal with whatever situation in which the BG globally and slowly changes in appearance. Therefore, the problem of *time of day* can generally be solved by these kind of methods. Algorithms assuming multimodal background models face the situation where the background appearance oscillates between two or more color ranges. This is particularly useful in dealing with outdoor situations where there are several moving parts in the scene or flickering areas, such as the tree leafs, flags, fountains, and sea surface. This situation is wellportrayed by the waving tree key problem. The other problems represent situations which imply in principle strong spatial reasoning, thus requiring per-region approaches. Let us discuss each of the problems separately: for each problem, we specify those approaches that explicitly focus on that issue.

Moved Objects. All the approaches examined fails in dealing with this problem, in the sense that an object moved in the scene, belonging to the scene, is detected as foreground for a certain amount of time. This amount depends on the adaptivity rate of the background model, that is, the faster the rate, the smaller the time interval.

Time of Day. BG model adaptivity ensures success in dealing with this problem, and almost each approach considered is able to solve it.

Global Light Switch. This problem is solved by those approaches which consider the global aspect of the scene. The main idea is that when a global change does occur in the scene, that is, when a consistent portion of the frame labeled as BG suddenly changes, a recovery mechanism is instantiated which evaluates the change as a sudden evolution of the BG model, so that the amount of false positive alarms re likely minimized. The techniques which explicitly deal with this problem are [15, 58, 59, 61, 65]. In all the other adaptive approaches, this problem generates a massive amount of false positives until when the learning rate "absorb" the novel aspect of the scene. Another solution consists in considering texture or edge information [36].

Local Light Switch. This problem is solved by those approaches which learn in advance how the illumination can locally change the aspect of the scene. Nowadays, the only approach which deals with this problem is [61].

Waving Trees. This problem is successfully faced by two classes of approaches. One is the per-pixel methods that admit a multimodal BG model (the movement of the tree is usually repetitive and holds for a long time, causing a multimodal BG). The other class is composed by the per-region techniques which inspect the neighborhood of a "source" pixel, looking whether the object portrayed in the source has locally moved or not.

Camouflage. Solving the camouflage issue is possible when other information other than the sole chromatic aspect is taken into account. For example, texture information greatly improves the BG subtraction [36, 37, 39]. The other source of information comes from the knowledge of the foreground; for example, employing contour information or connected-component analysis on the foreground, it is possible to recover the camouflage problem by performing morphological operations [15, 65].

Foreground Aperture. Even in this case, texture information improves the expressivity in the BG model, helping where the mere chromatic information leads to ambiguity between BG and FG appearances [36, 37, 39].

Sleeping Foreground. This problem is the most related with the FG modeling: actually, using only visual information and without having an exact knowledge of the FG appearance (which may help in detecting a still FG object which must remain separated from the scene), this problem cannot be solved. This is implied by the basic definition of the BG, that is, whatever visual static element and whose appearance does not change over time is, background.

Shadows. This problem can be faced employing two strategies: the first implies a per-pixel color analysis, which aims at modeling the range of variations assumed by the BG pixel values when affected by shadows, thus avoiding false positives. The most known approach in this class is [25], where the shadow analysis holds in the HSV color space. Other approaches try to define shadow-invariant color spaces [30, 32, 65]. The other class of strategies considers edge information, that is, more robust against shadows [36, 39, 40].

Reflections. This problem has been never considered in scenarios employing a single monocular video camera.

In general, the approaches that face simultaneously and successfully with several of the above problems (i.e., that present results on several Wallflower sequences) are [15, 36, 65].

4. Multiple Video Sensors

The majority of background subtraction techniques are designed for being used in a monocular camera framework which is highly effective for many common surveillance scenarios. Anyway, this setting encounters difficulties in dealing with sudden illumination changes, reflections, and shadows.

The use of two or more cameras for background modeling serves to overcome these problems. Illumination changes and reflections depend on the field of view of the camera and can be managed observing the scene from different view points, while shadows can be filtered out if 3D information is available. Even if it is possible to determine the 3D world positions of the objects in the scene with a single camera (e.g., [72]), this is in general very difficult and unreliable [73].

Therefore multicamera approaches to retrieve 3D information have been proposed, based on the following.

(i) *Stereo Camera.* A single device integrating two or more monocular cameras with small baseline (i.e., the distance between focal center of the cameras).

(ii) *Multiple Cameras.* A network of calibrated monocular or stereo cameras monitoring the scene from significantly different viewpoints.

4.1. Stereo Cameras. The disparity map extracted that correlates the two views of a stereo camera can be used as an input for a disparity-based background subtraction algorithm. In order to accurately model the background, a dense disparity map needs to be computed.

For obtaining an accurate dense map of correlations between two stereo images, time-consuming stereo algorithms are usually required. Without the aid of specialized hardware, most of these algorithms perform too slowly for real time background subtraction [74, 75]. As a consequence, state-of-the-art dedicated hardware solutions implement simple and less accurate stereo correlations methods instead of more precise ones [76]. In some cases, the correlation between left and right images is unreliable, and the disparity map presents holes due to "invalid" pixels (i.e., points with invalid depth values).

Stereo vision has been used in [77] to build the occupancy map of the ground plane as background model, that is, used to determine moving objects in the scene. The background disparity image is computed by averaging the stereo results from an initial background learning stage where the scene is assumed to contain no people. Pixels that have a disparity larger than the background (i.e., closer to the camera) are marked as foreground.

In [78], a simple bimodal model (normal distribution plus an unmodeled token) is used to build the background model. A similar approach is exploited in [79], where a histogram of disparity values across a range of time and gain conditions is computed. Gathering background observations over long-term sequences has the advantage that lighting variation can be included in the background training set. If background subtraction methods are based on depth alone [78, 80], errors due to foreground objects in close proximity to the background or foreground objects having homogeneous texture arise. The integration of color and depth information reduces the effect of the following problems:

(1) points with similar color background and foreground

(2) shadows

(3) invalid pixels in background or foreground

(4) points with similar depth in both background and foreground.

In [81], an example of a joint (color + depth) background estimation is given. The background model is based on a multidimensional (depth and RGB colors) histogram approximating a mixture of Gaussians, while foreground extraction is performed via background comparison in depth and normalized color.

In [82], a method for modeling the background that uses per-pixel, time-adaptive, Gaussian mixtures in the combined input space of depth and luminance-invariant color is proposed. The background model learning rate is modulated on the scene activity and the color-based segmentation criteria are dependent on depth observations. The method explicitly deals with illumination changes, shadows, reflections, camouflage, and changes in the background.

The same idea of integrating depth information and color intensity coming from the left view of the stereo sensor is exploited by the PLT system in [73]. It is a real-time system, based on a calibrated fixed stereo vision sensor. The system analyses three interconnected representations of the stereo data to dynamically update a model of the background, to extract foreground objects, such as people and rearranged furniture, and to track their positions in the world. The background model is a composition of intensity, disparity and edge information, and it is adaptively updated with a learning factor that varies over time and is different for each pixel.

4.2. Network of Cameras. In order to monitor large areas and/or managing occlusions, the only solution is to use multiple cameras. It is not straightforward to generalize a single-camera system to become a multicamera one, because of a series of problems like camera installation, camera calibration, object matching, and data fusion.

Redundant cameras increase not only processing time and algorithmic complexity, but also the installation cost. In contrast, a lack of cameras may cause some blind spots, that reduce the reliability of the surveillance system. Moreover, calibration is more complex when multiple cameras are employed and object matching among multiple cameras involves finding the correspondences between the objects in different images.

In [83], a real time 3D tracking system using three calibrated cameras to locate and track objects and people in a conference room is presented. A background model is computed for each camera view, using a mixture of Gaussians to estimate the background color per pixel. The background subtraction is performed on both the YUV and the RG color spaces. Matching RG foreground regions and YUV regions, is possible to cut off most of the shadows, thanks to the use of chromatic information, and, at the same time, to exploit intensity information to obtain smoother silhouettes.

M_2Tracker [84] uses a region-based stereo algorithm to find 3D points inside an object, and Bayesian Classification to classify each pixel as belonging to a person or the background. Taking into account models of the foreground objects in the scene, in addition to information about the background, leads to better background subtraction results.

In [85], a planar homography-based method combines foreground likelihood information (probability of a pixel in the image belonging to the foreground) from different views to resolve occlusions and determine the locations of people on the ground plane. The foreground likelihood maps in each view is estimated by modeling the background using a mixture of Gaussians. The approach fails in presence of strong shadows. Carnegie Mellon University developed a system [86] that allows a human operator to monitor activities over a large area using a distributed network of active video sensors. Their system can detect and track people and vehicles within cluttered scenes and monitor their activities over long periods of time. They developed robust routines for detecting moving objects using a combination of temporal differencing and template tracking.

EasyLiving project [87] aims to create a practical person-tracking system that solves most of the real-world problems. It uses two sets of color stereo cameras for tracking people during live demonstrations in a living room. Colour histograms are created for each detected person and are used to identify and track multiple people standing, walking, sitting, occluding, and entering or leaving the space. The background is modeled by computing the mean and variance for each pixel in the depth and color images over a sequence of 30 frames on the empty room.

In [74], a two-camera configuration is described, in which the cameras are vertically aligned with respect to a dominant ground plane (i.e., the baseline is orthogonal to the plane on which foreground objects appear). Background subtraction is performed by computing the normalized color difference for a background conjugate pair and averaging the component differences over a 3×3 neighborhood. Each background conjugate pair is modeled with a mixture of Gaussians. Foreground pixels are then detected if the associated normalized color differences fall outside a decision surface defined by a global false alarm rate.

4.3. Capabilities of the Approaches Based on Multiple Visual Sensors. The use of a stereo camera represent a compact solution, relatively cheap and easy to calibrate and set up, able to manage shadows and illumination changes. Indeed, the disparities information is more invariable to illumination changes with respect to the information provided by a single camera [88], and the insensitivity of stereo to changes in lighting mitigates to some extent the need for adaptation [77]. On the other hand, a multiple camera network allows to view the scene from many directions, monitoring an area larger than what a single stereo sensor can do. However, multicamera systems have to deal with problems in establishing geometric relationships between views and in maintaining temporal synchronization of frames.

In the following, we analyze those problems, taken from Section 2, for which the multiple visual sensor contribute in reaching optimal solutions.

Camouflage. This problem is effectively faced by integrating the depth information to the color information [73, 81, 82].

Foreground Aperture. Even in this case, texture information improves the expressivity in the BG model, helping where the mere chromatic information leads to ambiguity between the BG and the FG appearance [36, 37, 39].

Shadows. This issue is solved employing both stereo cameras [73, 81, 82] and camera networks [74, 83].

Reflections. The use of multiple camera permits to solve this problem: the solution is based on the 3D structure of the scene monitored. The 3D map permits to locate the ground plane of the scene, thus, to suppress all the specularities as those objects lying below this plane [74].

5. Single Audio Monaural Sensor

Analogously to image background modeling for video analysis, a logical initial phase in applying audio analysis to surveillance and monitoring applications is the detection of background audio. This would be useful to highlight sections of interest in an audio signal, like for example the sound of breaking glass.

There are a number of differences between the visual and audio domains, with respect to the data. The reduced amount of data in audio results in lower processing overheads, and encourages a more complex computational approach to analysis. Moreover, the characteristics of the audio usually exhibit a higher degree of variability. This is due to both the superimposition of multiple audio sources within a single input signal and the superimposition of the same sound at different times (multipatch echoing). Similar situations for video could occur through reflection off partially reflective surfaces. This results in the formation of complex and dynamic audio backgrounds.

Background audio can be defined as the recurring and persistent audio characteristics that dominates the portion of the signal. Foreground sounds detection can be carried out as the departure from this BG model.

Outside the automated surveillance context, several approaches to computational audio analysis are present, mainly focused on the computational translation of psychoacoustics results. One class of approaches is the so called *computational auditory scene analysis* (CASA) [89], aimed at the separation and classification of sounds present in

a specific environment. Closely related to this field there is the *computational auditory scene recognition* (CASR) [90, 91], aimed at an overall environment interpretation instead of analyzing the different sound sources. Besides various psychoacoustically oriented approaches derived from these two classes, a third approach, used both in CASA and CASR contexts, tried to fuse "blind" statistical knowledge with biologically driven representations of the two previous fields, performing audio classification and segmentation tasks [92], and source separation [93, 94] (i.e., blind source separation). In this last approach, many efforts are addressed to the speech processing area, in which the goal is to separate the different voices composing the audio pattern using several microphones [94] or only one monaural sensor [93].

In the surveillance context, some proposed methods in the field of BG subtraction are mainly based on the monitoring of the audio intensity [95–97], or are aimed at recognizing specific class of sounds [98]. These methods are not adaptive to the several possible audio situations, and they do not exploit all the potential information conveyed by the audio channel.

The following approaches, instead, are more general, they are adaptive and they can cope with quite complex backgrounds. In [99], the authors implement a version of the Gaussian Mixture Model (GMM) method in the audio domain. The audio signal, acquired by a single microphone, is processed by considering its frequency spectrum: it is subdivided in suitable subbands, assumed to convey independent information about the audio events. Each subband is modeled by a mixture of Gaussians. Being the model online updated over time, this makes the method adaptive to the possible different background situations. At each instant t, FG information is detected by considering the set of subbands that show atypical behaviors.

In [100], the authors also employ an online, unsupervised and adaptive GMM to model the states of the audio signal. Besides, they propose some solutions to more accurately model complex backgrounds. One is an entropy-based approach for combining fragmented BG models to determine the BG states of the signal. Then, the number of states to be incorporated into the background model is adaptively adjusted according to the background complexity. Finally, an auxiliary cache is employed, with to scope to prevent the removal from the system of potentially useful observed distributions when the audio is rapidly changing.

An issue not addressed by the previous methods, quite similar to the *Sleeping foreground* problem in video analysis (see below in Section 5.1), is when the foreground is gradual and longer lasting, like a plan passing overhead. If there is no a priori knowledge of the FG and BG, the system adapts the FG sound as background. This particular situation is addressed in [101], by incorporating explicit knowledge of data into the process. The framework is composed by two models. First, the models for the BG and FG sounds are learnt, using a semisupervised method. Then, the learned models are used to bootstrap the system. A separate model detects the changes in the background, and it is finally integrated with the audio predictions models to decide on the final FG/BG determination.

5.1. Capabilities of the Approaches Based on a Single Audio Sensor. The definition of audio background and its modelling for background subtraction incorporates issues that are analogous to those of the visual domain. In the following, we will consider the problems reported in Section 2, analyzing how they translate into the audio domain, and how they are solved by the nowadays approaches. Moreover, once a correspondence is found, we will define a novel name for an audio key issue, in order to gain in clarity.

In general, whereas the visual domain may be considered as formed by several independent entities, that is, the pixels signals, in the audio domain the spectral subband assume the meaning of the basic independent entities. This analogy is the one mostly used in the literature, and it will drive us in linking the different key problems across modalities.

Moved Object. This situation originally consists in a portion of the visual scene that is, moved. In the audio domain, a portion consists in an audio subband. Therefore, whatever approach that allows a local adaptation of the audio spectrum related to the BG solves this problem. The adaptation depends also in this case by a learning rate. The higher the rate, the faster the model adaptation [99, 100]. We will name this audio problem as *Local change*.

Time of Day. This problem shows in the audio when the BG spectrum slowly changes. Therefore, approaches that develop an adaptive model solve this problem [99, 100]. We will name this audio problem as *Slow evolution*.

Global Light Switch. Global light switch can be intended in the audio as an abrupt global change of the audio spectrum. In the video, a global change of illumination has not to be intended as a FG entity, because the change is global and persistent and because the structure of the scene does not change. The structure invariance in the video can be evaluated by employing edge or texture features, while it is not clear neither what is the structure of a environmental audio background, nor what are the features to model it. Therefore, an abrupt change in the audio spectrum will be evaluated as an evident presence of foreground and successively absorbed as BG if the BG model is adaptive, unless a classification-based approach is employed [99, 100], that minimizes the amount of FG by choosing the most suitable BG model across a set of BG models [101]. We will name this audio problem as *Global fast variation*.

Waving Trees. In audio, the analog of the waving tree problem is that of a multimodal audio background, in the sense that each independent entity of the model, that is, the audio subband, shows a multimodal statistics. This happens for example when repeated signals occurs in the scene (the sound produced by a factory machine). Therefore, approaches that deal with multimodality (as expressed above) in the BG modelling deal with this problem successfully [99, 100]. We will name this audio problem as *Repeated background*.

Camouflage. The camouflage in the audio can be reasonably seen as the presence of a FG sound which is similar to that of the BG. Using the audio spectrum as basic model for the BG characterization solves the problem of camouflage, because different sounds having the same spectral characteristic (so, when we are in presence of similar sounds) will produce a spectrum where the spectral intensities are summed over. Such spectrum is different to that of the single BG sound, where the intensities are lower. We will name this audio problem as *Audio camouflage*.

Sleeping Foreground. The sleeping foreground occurs in the audio when a FG sound continuously holds, becoming BG. This issue may be solved explicitly by employing FG models, as done in [101]. We will name this audio problem as *Sleeping audio foreground*.

It is worth noting that in this case, the visual problems of Local light switch, Foreground aperture, Shadows and Reflections have not a clear correspondence in the audio domain, and thus they are omitted from the analysis.

6. Single Infrared Sensor

Most algorithms for object detection are designed only for daytime visual surveillance and are generally not effective for dealing with night conditions, when the images have low brightness, low contrast, low signal-to-noise ratio (SNR) and nearly no color information [102].

For night-vision surveillance, two primary technologies are used: image enhancement and thermal imaging.

Image enhancement techniques aim to amplify the light reflected by the objects in the monitored scene to improve visibility. Infrared (IR) light levels are high at twilight or in halogen light, therefore a camera with good IR sensitivity can capture short-wavelength infrared (SWIR) emissions to increase the image quality. SWIR wavelength follows directly from the visible spectrum (VIS), and therefore it is also called near infrared.

Thermal imaging refers to the process of capturing the long-wave IR radiation emitted or reflected by objects in the scene, which is undetectable to the human eye, and transforming it into a colored or grayscale image.

The use of infrared light and night vision devices should not be confused with thermal imaging (see Figure 5 for a visual comparison). If scene is completely dark, then image enhancement methods are not effective and it is necessary to use a thermal infrared camera. However, the cost of a thermal camera is too high for most surveillance applications.

6.1. Near Infrared Sensors. Near infrared (NIR) sensors are low cost (around 100 dollars) when compared with thermal infrared sensors (around 1000 dollars) and have a much higher resolution. NIR cameras are suitable for environments with a low illumination level, typically between 5 and 50 *lux* [103]. In urban surveillance, it is not unusual to have artificial light sources illuminating the scene at night (e.g., monitored parking lots next to buildings tends to be well lit). NIR sensors represent a cheaper alternative to thermal cameras for monitoring these urban scenarios. However, SWIR-based video surveillance presents a series of challenges [103].

(i) *Low SNR.* With low light levels, a high gain is required to enhance the image brightness. However, a high gain tends to amplify the sensor's noise introducing a considerable variance in pixel intensity between frames that impairs the background modeling approaches based on statistical analysis.

(ii) *Blooming.* The presence of strong light sources (e.g. car headlights and street lamps) can lead to the saturation of the pixel involved, deforming the detected shape of objects.

(iii) *Reflections.* Surfaces in the scene can reflect light causing false positives.

(iv) *Shadows.* Moving objects cause sharp shadows with changing orientation (with respect to the object).

In [103], a system to perform automated parking lot surveillance at night time is presented. As a preprocessing step, contrast and brightness of input images are enhanced and spatial smoothing is applied. The background model is built as a mixture of Gaussians. In [104], an algorithm for background modeling based on spatiotemporal patches especially suited for night outdoor scenes is presented. Based on the spatiotemporal patches, called bricks, the background models are learned by an on-line subspace learning method. However, the authors claim the algorithm fails on surfaces with specular reflection.

6.2. Thermal Infrared Sensors. Thermal infrared sensors (see Figure 6) are not subject to color imagery problems in managing shadows, sudden illumination changes, and poor night-time visibility. However, thermal imagery has to deal with its own particular challenges.

(i) Commonly used ferroelectric BST thermal sensor yields imagery with a low SNR, which results in limited information for performing detection or tracking tasks.

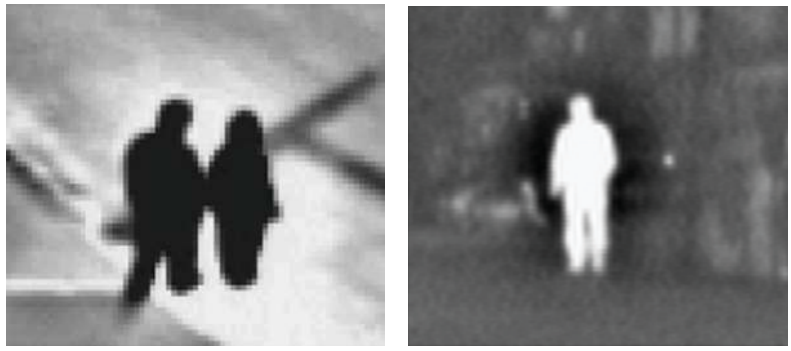
(ii) Uncalibrated polarity and intensity of the thermal image, that is, the disparity in terms of thermal properties between the foreground and the background is quite different if the background is warm or cold (see Figure 7).

(iii) Saturation or "halo effect", that appears around very hot or cold objects, can modify the geometrical properties of the foreground objects deforming their shape.

The majority of the object detection algorithms working with the thermal domain adopt a simple thresholding method to build the foreground mask, assuming that a foreground object is much hotter than the background and hence appears brighter, as an "hot-spot" [105]. In [106], a thresholded image is computed as the first step of a human posture estimation method, based on the assumption

(a) (b)

Figure 6: A color image (a) and a thermal image (b) from OSU Color-Thermal Database [17, 105].



(a) (b)

Figure 7: Uncalibrated polarity and halo issues in thermal imagery: (a) bright halo around dark objects [105], (b) dark halo around bright object [110].

that the temperature of the human body is hotter than the background. The hot-spot assumption is used in [107] for developing an automatic gait recognition method where the silhouettes are extracted by thresholding. In [108], the detection of hotspots is performed using a flexible threshold calculated as the balance between the thermal image mean intensity and the highest intensity, then a Support Vector Machines-(SVM-) based approach aims to classify humans. In [109] the threshold value is extracted from a training dataset of rectangular boxes containing pedestrians, then probabilistic templates are exploited to capture the variations in human shape, for managing the case where contrast is low and body parts are missing.

However, the hot-spot assumption does not hold if the scene is monitored in different time of the day and/or at different environmental temperatures (e.g., during winter or summer). Indeed, in night-time (or during winter) usually, foreground is warmer than background, but this is not always true in day-time (or summer), when the background can be warmer than the foreground.

Moreover, the presence of halos in thermal imagery compromises the use of traditional visual background subtraction techniques [105]. Since the halo surrounding the moving object usually diverges from the background model, it is classified as foreground introducing an error in retrieving the structural properties of the foreground objects.

The above discussed challenges in using thermal imagery have been largely ignored in the past [105]. Integrating visual and thermal imagery can lead to overcome those drawbacks. Indeed, in presence of sufficient illumination conditions, colour optical sensors are oblivious to temperature differences in the scene and are typically more effective than thermal cameras when the thermal properties of the objects in the scene are similar to the surrounding environment.

6.3. Capabilities of the Approaches Based on a Single Infrared Sensor. Taken alone and evaluated in scenarios where the illumination is enough to perform also visual background subtraction, infrared sensory cannot provide robust systems for the background subtraction, for all the limits discussed above. Anyway, infrared is effective when the illumination is scarce, and in disambiguating a camouflage situation, where the visual aspect of the FG is similar to that of the BG. Infrared is also the only working solution in scenarios where the FG objects lie on water surfaces, since the false positive detections caused by waves can be totally filtered out.

7. Fusion of Multiple Sensors

One of the most desirable qualities of a video surveillance system is *persistence*, or the ability to be effective all the times. However, a single sensor is generally not effective

in all situations. The use of complementary sensors, hence, becomes important to provide complete and sufficient information: information redundancy permits to validate observations, in order to enhance FG/BG separation, and it becomes essential when one modality is not available.

Fusing data from heterogeneous information sources arises new problems, such as how to associate distinct objects that represent the same entity. Moreover, the complexity of the problem increases when the sources do not have a complete knowledge about the monitoring area and in situations where the sensors measurements are ambiguous and imprecise.

There is an increasing interest in developing multimodal systems that can simultaneously analyze information from multiple sources of information. The most interesting trends regard the fusion of thermal and visible imagery and the fusion of audio and video information.

7.1. Fusion of Thermal and Visible Imagery. Thermal and color video cameras are both widely used for surveillance. Thermal cameras are independent of illumination, so they are more effective than color cameras under poor lighting conditions. On the other hand, color optical sensors does not consider temperature differences in the scene, and are typically more effective than thermal cameras when the thermal properties of the objects in the scene are similar to the surrounding environment (provided that the scene is well illuminated and the objects have color signatures different from the background). Integrating visual and thermal imagery can lead to overcome the drawback of both sensors, enhancing the overall performance (Figure 8).

In [105], a three-stage algorithm to detect the moving objects in urban settings is described. Background subtraction is performed on thermal images, detecting the regions of interest in the scene. Color and intensity information is used within these areas to obtain the corresponding regions of interest in the visible domain. Within each image region (thermal and visible, treated independently) the input and background gradient information are combined as to highlight only the contours of the foreground object. Contour fragments belonging to corresponding region in the thermal and visible domains are then fused, using the combined input gradient information from both sensors. This technique permits to filter out both halos and shadows. A similar approach that uses gradient information from both visible and thermal images is described in [112]: the fusion step is based on mutual agreement between the two modalities. In [113], the authors propose to use a IR camera in conjunction with a standard camera for detecting humans. Background subtraction is performed independently on both camera images using a single Gaussian probability distribution to model each background pixel. The couple of detected foreground masks is extracted using a hierarchical genetic algorithm, and the two registered silhouettes are then fused together into the final estimate. Another similar approach for humans detection is described in [111]. Even in this case BG subtraction is run on the two cameras independently, extracting the blobs from each camera. The blobs are then matched and aligned to reject false positives.

In [114], instead, an image fusion scheme that employs multiple scales is illustrated. The method first computes pixel saliency in the two images (IR and visible) at multiple scales, then a merging process, based on a measure of the difference in brightness across the images, produces the final foreground mask.

7.1.1. Capabilities of the Approaches Based on the Fusion of Thermal and Visible Imagery. In general, thermal imagery is taken as support for the visual modality. Considering the literature, the key problem in Section 2 where the fusion of thermal and visible imagery results particularly effective is that of the shadows: actually, all the approaches stress this fact in their experimental sections.

7.2. Fusion of Audio and Video Information. Many researchers have attempted to integrate vision and acoustic senses, with the aim to enhance object detection and tracking, more than BG subtraction. The typical scenario in an indoor environment with moving or static objects that produce sounds, monitored with fixed or moving cameras and fixed acoustic sensors.

For completeness we report in the following some of these methods, even if they do not tackle BG subtraction explicitly. Usually each sense is processed separately and the overall results are integrated in the final step. The system developed in [115], for example, uses an array of eight microphones to initially locate a speaker and then steer a camera towards the sound source. The camera does not participate in the localization of objects, but it is used to take images of the sound source after it has been localized. However, in [116], the authors demonstrate that the localization integrating audio and video information is more robust compared to the localization based on stand alone microphone arrays. In [117], the authors detect walking persons, with a method based on video sequences and step sounds. The audiovisual correlation is learned by a time-delay neural network, which then performs a spatiotemporal search for the walking person. In [118], the authors propose a quite complete surveillance system, focused on the integration of the visual and the audio information provided by different sensing agents. Static cameras, fixed microphones and mobile vision agents work together to detect intruders and to capture a closed image of them. In [119], the authors deal with tracking and identifying multiple people using discriminative visual and acoustic features extracted from cameras and microphone array measurements. The audio local sensor performs sound sources localization and source separation to extract the existing speeches in the environment; the video local sensor performs people localization and face-color extraction. The association decision is based on the belief theory, and the system provides robust performances even with noisy data.

A paper that instead focuses on fusing video and acoustic signals with the aim to enhance BG modeling is [120]. The authors build a multimodal model of the scene background, in which both the audio and the video are modeled by

(a) (b) (c)

Figure 8: Example of fusion of video and thermal imagery: (a), FG obtained from the thermal camera; at the center, FG obtained from the video camera; (b), their fusion result [111].

employing a time-adaptive mixture model. The system is able to detect single auditory or visual events, as well as audiovideo simultaneous situations, considering a synchrony principle. This integration permits to address the FG sleeping problem: an audiovisual pattern can remain an actual foreground even if one of the components (audio or video) becomes BG. The setting is composed by one fixed camera and a single microphone.

7.2.1. Capabilities of the Approaches Based on the Fusion of Audio and Video Information. Coupling the audio and the visual signal is a novel direction for the background subtraction literature. Actually, most of the approaches presented in the previous section propose a coupled modeling for the foreground, instead of detailing a pure background subtraction strategy. Anyway, all those approaches work in a clear setting, that is, where the audio signal is clearly associated to the foreground entities. Therefore, the application of such techniques in real-world situations need to be supported by technique able to perform the subtraction of useless information in both the audio and the visual channels. In this sense, [120] is the approach that more leads in this direction (even if it also proposes a modeling for the foreground entities).

8. How the Key Problems of Background Subtraction May Be Solved?

In this paper, we examined different approaches for the background subtraction, with a particular attention to how they solve typical hoary issues. We consider different sensor channels, and different multichannel integration policies. In this section we consider together all these techniques, summarizing for each problem what are the main strategies adopted to solve it.

In particular, we focus in the problems presented in Section 2, without considering the translated versions of the problems in the audio channel (Section 5.1). The table in Table 1 summarizes the main categories of methods described in this paper, and the problems that they explicitly solve.

Moreover, we individuate those that could be winning strategies that have not been completely exploited in the literature, hoping that some of them could be embraced and applied satisfactorily.

Moved Object (MO). In this case, mainly visual approaches are present in the literature, which are not able to solve this issue satisfactorily. Actually, when an object belonging to the scene is moved, it erroneously appears to be a FG entity, until when the BG model adapts and absorbs the novel visual layout. A useful direction to solve effectively this issue is considering thermal information: actually, if the background has thermal characteristics that are different from the FG objects, the visual change provoked by an object which is relocated may be inhibited by its thermal information.

Time of Day (TD). Adaptive BG models showed to be effective to definitely solve this issue. When the illumination is very scarce, thermal imagery may help. A good direction could be building a structured model that introduces the thermal imagery selectively, in order to maximize the BG/FG discrimination.

Light Switch (LS). This problem has been considered under a pure visual sense. The solutions present in the literature are satisfying, and operate by considering the global appearance of the scene. When a global abrupt change happens, the BG model is suddenly adapted or selected from a set of predetermined models, in order to minimize the amount of false positive alarms.

Local Light Switch (LLS). Local light switch is a novel problem, introduced here and scarcely considered in the literature. The approaches that face this problems work on the visual channel, studying in a bootstrap phase how the illumination of the scene locally changes, monitoring when a local change does occur and adapting the model consequently.

Waving Trees (WT). The oscillation of the background is effectively solved in the literature under a visual perspective. The idea is that the BG models have to be multimodal: this works well especially when the oscillation of the background

(or part of it) is persistent and well located (i.e., the oscillation has to occur for a long time in the same area; in other words, it has to be predictable). When the oscillations are rare or unpredictable, approaches that consider per-region strategies are decisive. The idea is that per-pixel models share their parameters, so that a background value in a pixel may be evaluated as BG even if it occurs in a local neighborhood.

Camouflage (C). Camouflage effects derive from the similarity between the features that characterize the foreground and those used for modeling the background. Therefore, the more discriminating features, the better the separation between FG and BG entities. In this case, under a visual perspective, gray level is the worst solution as feature. Moving to color values offers a better discriminability, that can be further ameliorated by employing edge and texture information. Particularly effective is the employment of stereo sensors, that introduce depth information in the analysis. Again, thermal imagery may help. A mixing of visual and thermal channels exploiting stereo devices has been never taken into account, and seems to be a reasonable novel strategy.

Bootstrapping (B). Bootstrapping methods are explicitly faced only under a visual perspective, by approaches of background initialization. These approaches offer good solutions: they essentially build statistics for devising a BG model by exploiting the principle of temporal persistence (elements of the scene which appear continuously with the same layout represent the BG) and spatial continuity (i.e., homogeneously colored surfaces or portions of the scene which exploit edge continuity belong to the BG). Bootstrapping considering other sensor channels has never been taken into account.

Foreground Aperture (FA). The problem of the spatiotemporal persistence of a foreground object, and its partial erroneous absorption in the BG model, has been faced in the literature under the sole visual modality. This problem primarily depends on a too fast learning rate of the BG model. Resolutive approaches employ per-region reasoning, by examining the detected FG regions and looking for holes, filling them by morphological operators. Foreground aperture considering other sensor channels has never been taken into account.

Sleeping Foreground (SF). This problem is the one that more implies a sort of knowledge of the FG entities, crossing the border towards goals that are typical of the tracking literature. In practice, the intuitive solution for this problem consists to inhibit the absorption mechanism of the BG model whereas a FG object occurs in the scene. In the literature, a solution comes through the use of multiple sensor channels. Employing thermal imagery associated to visual information permits to discriminate between FG and BG in an effective way. Actually, the background is assumed to be at a different temperature with respect to the FG objects: this contrast has to be maintained over time, so a still foreground will be always differentiated from the background. Employing audio signals is another way. Associating an audio pattern to a FG entity permits to enlarge the set of features that need to be constant in time for provoking a total BG absorption. Therefore, a visual entity (a person) which is still, that however maintains FG audio characteristics (i.e., that of being unexpected) remains a FG entity. Employing multiple sensor channels allows to solve this problem without relying on tracking techniques: that is, the idea is to enrich the BG model, in order to detect better FG entities, that is, entities that diverge from that model.

Shadows (SH). The solution for the shadows problem comes from the visual domain or employing multiple sensors or considering thermal imagery. In the first way, color analysis is applied, by building a chromatic range over which a background color may vary when affected by shadows. Otherwise, edge, or texture analysis, that has been shown to be robust to shadows, is applied. Stereo sensors discard the shadows simply relying on depth information, and multiple cameras are useful to build a 3D map where the items that are projected on the ground plane of the scene are labelled as shadows. Thermal imagery is oblivious to shadows issues.

Reflections (R). Reflections is a brand-new problem for the background subtraction literature, in the sense that very few approaches have been focused on this issue. It is more difficult than dealing with the shadows, because, as visible in our test sequence, reflections carry color, edge, or texture information which is not brought by shadows. Therefore, methods that rely on color, edge, and texture analysis fail. The only satisfying solution comes through the use of multiple sensors. A 3D map of the scene can be built (so, the BG model is enriched and made more expressive) and geometric assumptions on where a FG object could appear or not help in discarding reflection artifacts. The use of thermal imagery and stereo sensor is intuitively useful to solve this problem, but in the literature there are not approaches that explicitly deal with this problematic.

9. Final Remarks

In this paper, we present an essay of background subtraction methods. It has two important characteristics that make it diverse and appealing with respect to the other reviews. First, it considers different sensor channels and various integration policies of heterogeneous channels with which background subtraction may be carried out. This has never appeared before in the literature. Second, it is problem-oriented, that is, it individuates the key problems for the background subtraction and we analyze and discuss how the different approaches behave with respect to them. This permits to synthesize a global snapshot of the effectiveness of the nowadays background subtraction approaches. Almost each problem analyzed has a proper solution, that comes from different modalities or multimodal integration policies. Therefore, we hope that this problem-driven analysis may serve in devising an even more complete background subtraction system, able

to join sensor channels in an advantageous way, facing all the problems at the same time and providing convincing performances.

Acknowledgments

This paper is founded by the EU-Project FP7 SAMURAI, Grant FP7-SEC- 2007-01 no. 217899.

References

[1] H. T. Nguyen and A. W. M. Smeulders, "Robust tracking using foreground-background texture discrimination," *International Journal of Computer Vision*, vol. 69, no. 3, pp. 277–293, 2006.

[2] R. T. Collins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1631–1643, 2005.

[3] H.-T. Chen, T.-L. Liu, and C.-S. Fuh, "Probabilistic tracking with adaptive feature selection," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04)*, pp. 736–739, August 2004.

[4] F. Martez-Contreras, C. Orrite-Urunuela, E. Herrero-Jaraba, H. Ragheb, and S. A. Velastin, "Recognizing human actions using silhouette-based HMM," in *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS '09)*, pp. 43–48, 2009.

[5] H. Grabner and H. Bischof, "On-line boosting and vision," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 260–267, June 2006.

[6] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, June 2008.

[7] M. Bicego, M. Cristani, and V. Murino, "Unsupervised scene analysis: a hidden Markov model approach," *Computer Vision and Image Understanding*, vol. 102, no. 1, pp. 22–41, 2006.

[8] S. Gong, J. Ng, and J. Sherrah, "On the semantics of visual behaviour, structured events and trajectories of human action," *Image and Vision Computing*, vol. 20, no. 12, pp. 873–888, 2002.

[9] M. Piccardi, "Background subtraction techniques: a review," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '04)*, pp. 3099–3104, October 2004.

[10] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE Transactions on Image Processing*, vol. 14, no. 3, pp. 294–307, 2005.

[11] Y. Benezeth, P. M. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, "Review and evaluation of commonly-implemented background subtraction algorithms," in *Proceedings of the 19th International Conference on Pattern Recognition (ICPR '08)*, December 2008.

[12] A. Mittal and N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, vol. 2, pp. 302–309, 2004.

[13] S.-C. S. Cheung and C. Kamath, "Robust techniques for background subtraction in urban traffic video," in *Visual Communications and Image Processing*, vol. 5308 of *Proceedings of SPIE*, pp. 881–892, San Jose, Calif, USA, 2004.

[14] D. H. Parks and S. S. Fels, "Evaluation of background subtraction algorithms with post-processing," in *Proceedings of the 5th International Conference on Advanced Video and Signal Based Surveillance (AVSS '08)*, pp. 192–199, September 2008.

[15] WALLFLOWER, "Test images for wallflower paper," http://research.microsoft.com/en-us/um/people/jckrumm/wallflower/testimages.htm.

[16] "C. for Biometrics and S. Research. Cbsr nir face dataset," http://www.cbsr.ia.ac.cn .

[17] "OTCBVS Benchmark Dataset Collection," http://www.cse.ohio-state.edu/otcbvs-bench/.

[18] R. Miezianko, "Terravic research infrared database".

[19] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: principles and practice of background maintenance," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 255–261, 1999.

[20] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1778–1792, 2005.

[21] R. Jain and H. H. Nagel, "On the analysis of accumulative difference pictures from image sequences of real world scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 206–214, 1978.

[22] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[23] J. Heikkilä and O. Silven, "A real-time system for monitoring of cyclists and pedestrians," in *Proceedings of the 2nd IEEE International Workshop on Visual Surveillance*, pp. 74–81, Fort Collins, Colo, USA, 1999.

[24] N. J. B. McFarlane and C. P. Schofield, "Segmentation and tracking of piglets in images," *Machine Vision and Applications*, vol. 8, no. 3, pp. 187–193, 1995.

[25] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1337–1342, 2003.

[26] I. Haritaoglu, R. Cutler, D. Harwood, and L. S. Davis, "Backpack: detection of people carrying objects using silhouettes," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 385–397, 2001.

[27] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809–830, 2000.

[28] N. Friedman and S. Russell, "Image segmentation in video sequences: a probabilistic approach," in *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI '97)*, pp. 175–181, Morgan Kaufmann Publishers, San Francisco, Calif, USA, 1997.

[29] C. Stauffer and W. E.L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '99)*, vol. 2, pp. 246–252, 1999.

[30] H. Wang and D. Suter, "A re-evaluation of mixture-of-Gaussian background modeling," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, pp. 1017–1020, March 2005.

[31] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04)*, pp. 28–31, August 2004.

[32] A. Elgammal, D. Harwood, and L. Davis, "Non parametric model for background substraction," in *Proceedings of the 6th European Conference Computer Vision*, Dublin, Ireland, June-July 2000.

[33] A. M. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric model for background subtraction," in *Proceedings of the 6th European Conference on Computer Vision*, pp. 751–767, 2000.

[34] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley & Sons, New York, NY, USA, 2001.

[35] M. Levine, *Vision by Man and Machine*, McGraw-Hill, New York, NY, USA, 1985.

[36] P. Noriega and O. Bernier, "Real time illumination invariant background subtraction using local kernel histograms," in *Proceedings of the British Machine Vision Conference*, 2006.

[37] M. Heikkila and M. Pietikainen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 657–662, 2006.

[38] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Proceedings of the International Conference on Pattern Recognition (ICPR '94)*, pp. 582–585, 1994.

[39] J. Yao and J.-M. Odobez, "Multi-layer background subtraction based on color and texture," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, June 2007.

[40] B. Klare and S. Sarkar, "Background subtraction in varying illuminations using an ensemble based on an enlarged feature set," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 66–73, 2009.

[41] M. Cristani and V. Murino, "A spatial sampling mechanism for effective background subtraction," in *Proceedings of the 2nd International Conference on Computer Vision Theory and Applications (VISAPP '07)*, pp. 403–410, March 2007.

[42] O. Barnich and M. Van Droogenbroeck, "ViBE: a powerful random technique to estimate the background in video sequences," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '09)*, pp. 945–948, IEEE Computer Society, Washington, DC, USA, 2009.

[43] S. Rowe and A. Blake, "Statistical mosaics for tracking," *Image and Vision Computing*, vol. 14, no. 8, pp. 549–564, 1996.

[44] A. Mittal and D. Huttenlocher, "Scene modeling for wide area surveillance and image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '00)*, vol. 2, pp. 160–167, June 2000.

[45] E. Hayman and J.-O. Eklundh, "Statistical background subtraction for a mobile observer," in *Proceedings of the 9th IEEE International Conference on Computer Vision*, vol. 1, pp. 67–74, 2003.

[46] Y. Ren, C.-S. Chua, and Y.-K. Ho, "Statistical background modeling for non-stationary camera," *Pattern Recognition Letters*, vol. 24, no. 1–3, pp. 183–196, 2003.

[47] M. Irani and P. Anandan, "A unified approach to moving object detection in 2d and 3d scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 6, pp. 577–589, 1998.

[48] H. S. Sawhney, Y. Guo, and R. Kumar, "Independent motion detection in 3D scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1191–1199, 2000.

[49] C. Yuan, G. Medioni, J. Kang, and I. Cohen, "Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric constraints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1627–1641, 2007.

[50] H. Tao, H. S. Sawhney, and R. Kumar, "Object tracking with bayesian estimation of dynamic layer representations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 75–89, 2002.

[51] J. Xiao and M. Shah, "Motion layer extraction in the presence of occlusion using graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1644–1659, 2005.

[52] Y. Jin, L. Tao, H. Di, N. I. Rao, and G. Xu, "Background modeling from a free-moving camera by multi-layer homography algorithm," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '08)*, pp. 1572–1575, October 2008.

[53] R. Vidail and Y. Ma, "A unified algebraic approach to 2-D and 3-D motion segmentation," in *Proceedings of the 8th European Conference on Computer Vision*, vol. 3021 of *Lecture Notes in Computer Science*, pp. 1–15, Prague, Czech Republic, May 2004.

[54] K. Kanatani, "Motion segmentation by subspace separation and model selection," in *Proceedings of the 8th International Conference on Computer Vision*, vol. 2, pp. 586–591, July 2001.

[55] Y. Sheikh, O. Javed, and T. Kanade, "Background subtraction for freely moving cameras," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '09)*, pp. 1219–1225, 2009.

[56] R. T. Collins and Y. Liu, "On-line selection of discriminative tracking features," in *Proceedings of the 9th IEEE International Conference on Computer Vision*, vol. 1, pp. 346–352, October 2003.

[57] T. Parag, A. Elgammal, and A. Mittal, "A framework for feature selection for background subtraction," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 1916–1923, June 2006.

[58] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, and J. M. Buhmann, "Topology free hidden Markov models: application to background modeling," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 294–301, 2001.

[59] N. Ohta, "A statistical approach to background subtraction for surveillance systems," in *Proceedings of the 8th International Conference on Computer Vision*, vol. 2, pp. 481–486, July 2001.

[60] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–843, 2000.

[61] M. Cristani, M. Bicego, and V. Murino, "Integrated region- and pixel-based approach to background modelling," in *Proceedings of the IEEE Workshop on Motion and Video Computing*, pp. 3–8, 2002.

[62] M. Cristani, M. Bicego, and V. Murino, "Multi-level background initialization using hidden Markov models," in *Proceedings of the ACMSIGMM Workshop on Video Surveillance*, pp. 11–19, 2003.

[63] Q. Xiong and C. Jaynes, "Multi-resolution background modeling of dynamic scenes using weighted match filters," in *Proceedings of the 2nd ACM International Workshop on Video Sureveillance and Sensor Networks (VSSN '04)*, pp. 88–96, ACM Press, New York, NY, USA, 2004.

[64] J. Park, A. Tabb, and A. C. Kak, "Hierarchical data structure for real-time background subtraction," in *Proceedings of International Conference on Image Processing (ICIP '06)*, 2006.

[65] H. Wang and D. Suter, "Background subtraction based on a robust consensus method," in *Proceedings of International Conference on Pattern Recognition*, vol. 1, pp. 223–226, 2006.

[66] X. Gao, T. E. Boult, F. Coetzee, and V. Ramesh, "Error analysis of background adaption," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '00)*, vol. 1, pp. 503–510, June 2000.

[67] B. Gloyer, H. K. Aghajan, K. Siu, and T. Kailath, "Video-based freeway-monitoring system using recursive vehicle tracking," in *Image and Video Processing III*, vol. 2421 of *Proceedings of SPIE*, pp. 173–180, San Jose, Calif, USA, 1995.

[68] W. Long and Y.-H. Yang, "Stationary background generation: an alternative to the difference of two images," *Pattern Recognition*, vol. 23, no. 12, pp. 1351–1359, 1990.

[69] D. Gutchess, M. Trajkovicz, E. Cohen-Solal, D. Lyons, and A. K. Jain, "A background model initialization algorithm for video surveillance," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '01)*, vol. 1, pp. 733–740, 2001.

[70] A. Colombari, M. Cristani, V. Murino, and A. Fusiello, "Exemplarbased background model initialization," in *Proceedings of the third ACM International Workshop on Video Surveillance and Sensor Networks*, pp. 29–36, Hilton, Singapore, 2005.

[71] A. Colombari, A. Fusiello, and V. Murino, "Background initialization in cluttered sequences," in *Proceedings of the 5th Conference on Computer Vision and Pattern Recognition (CVPR '06)*, 2006.

[72] T. Zhao and R. Nevatia, "Tracking multiple humans in crowded environment," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, pp. 406–413, July 2004.

[73] S. Bahadori, L. Iocchi, G. R. Leone, D. Nardi, and L. Scozzafava, "Real-time people localization and tracking through fixed stereo vision," *Applied Intelligence*, vol. 26, no. 2, pp. 83–97, 2007.

[74] S. Lim, A. Mittal, L. Davis, and N. Paragios, "Fast illuminationinvariant background subtraction using two views: error analysis, sensor placement and applications," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 1071–1078, 2005.

[75] M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 993–1008, 2003.

[76] N. Lazaros, G. C. Sirakoulis, and A. Gasteratos, "Review of stereo vision algorithms: from software to hardware," *International Journal of Optomechatronics*, vol. 2, no. 4, pp. 435–462, 2008.

[77] D. Beymer and K. Konolige, "Real-time tracking of multiple people using continuous detection," in *Proceedings of International Conference on Computer Vision (ICCV '99)*, 1999.

[78] C. Eveland, K. Konolige, and R. C. Bolles, "Background modeling for segmentation of video-rate stereo sequences," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 266–271, June 1998.

[79] T. Darrell, D. Demirdjian, N. Checka, and P. Felzenszwalb, "Plan-view trajectory estimation with dense stereo background models," in *Proceedings of the 8th International Conference on Computer Vision (ICCV '01)*, pp. 628–635, July 2001.

[80] Y. Ivanov, A. Bobick, and J. Liu, "Fast lighting independent background subtraction," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 199–207, 2000.

[81] G. Gordon, T. Darrell, M. Harville, and J. Woodfill, "Background estimation and removal based on range and color," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '99)*, pp. 459–464, June 1999.

[82] M. Harville, G. Gordon, and J. Woodfill, "Foreground segmentation using adaptive mixture models in color and depth," in *Proceedings of the IEEE Workshop on Detection and Recognition of Events in Video*, pp. 3–11, 2001.

[83] D. Focken and R. Stiefelhagen, "Towards vision-based 3-d people tracking in a smart room," in *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, pp. 400–405, 2002.

[84] A. Mittal and L. S. Davis, "M2tracker: a multi-view approach to segmenting and tracking people in a cluttered scene," *International Journal of Computer Vision*, vol. 51, no. 3, pp. 189–203, 2003.

[85] S. M. Khan and M. Shah, "A multiview approach to tracking people in crowded scenes using a planar homography constraint," in *Proceedings of the 9th European Conference on Computer Vision (ECCV '06)*, vol. 3954 of *Lecture Notes in Computer Science*, pp. 133–146, Graz, Austria, 2006.

[86] A. J. Lipton, H. Fujiyoshi, and R. S. Patil, "Moving target classification and tracking from real-time video," in *Proceedings of the IEEE Workshop Application of Computer Vision*, pp. 8–14, 1998.

[87] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-camera multi-person tracking for easyliving," in *Proceedings of the 3rd IEEE International Workshop on Visual Surveillance (VS '00)*, p. 3, 2000.

[88] R. Muñoz-Salinas, E. Aguirre, and M. García-Silvente, "People detection and tracking using stereo vision and color," *Image and Vision Computing*, vol. 25, no. 6, pp. 995–1007, 2007.

[89] A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, London, UK, 1990.

[90] V. Peltonen, *Computational auditory scene recognition*, M.S. thesis, Tampere University of Tech., Tampere, Finland, 2001.

[91] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2895–2907, 2003.

[92] T. Zhang and C.-C. Jay Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 441–457, 2001.

[93] S. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems*, pp. 793–799, 2000.

[94] K. Hild II, D. Erdogmus, and J. Principe, "On-line minimum mutual information method for time-varying blind source separation," in *Proceedings of the International Workshop*

on Independent Component Analysis and Signal Separation (ICA '01), pp. 126–131, 2001.

[95] M. Stager, P. Lukowica, N. Perera, T. V. Buren, G. Troster, and T. Starner, "Soundbutton: design of a low power wearable audio classification system," in *Proceedings of the 7th IEEE International Symposium on Wearable Computers*, pp. 12–17, 2003.

[96] J. Chen, A. H. Kam, J. Zhang, N. Liu, and L. Shue, "Bathroom activity monitoring based on sound," in *Proceedings of the 3rd International Conference on Pervasive Computing*, vol. 3468 of *Lecture Notes in Computer Science*, pp. 47–61, Munich, Germany, May 2005.

[97] M. Azlan, I. Cartwright, N. Jones, T. Quirk, and G. West, "Multimodal monitoring of the aged in their own homes," in *Proceedings of the 3rd International Conference on Smart Homes and Health Telematics (ICOST '05)*, 2005.

[98] D. Ellis, "Detecting alarm sounds," in *Proceedings of Consistent and Reliable Acoustic Cues for sound analysis(CRAC '01)*, Aalborg, Denmark, September 2001.

[99] M. Cristani, M. Bicego, and V. Murino, "On-line adaptive background modelling for audio surveillance," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04)*, vol. 2, pp. 399–402, August 2004.

[100] S. Moncrieff, S. Venkatesh, and G. West, "Online audio background determination for complex audio environments," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 3, no. 2, Article ID 1230814, 2007.

[101] S. Chu, S. Narayanan, and C.-C. J. Kuo, "A semi-supervised learning approach to online audio background detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '09)*, pp. 1629–1632, April 2009.

[102] K. Huang, L. Wang, T. Tan, and S. Maybank, "A real-time object detecting and tracking system for outdoor night surveillance," *Pattern Recognition*, vol. 41, no. 1, pp. 432–444, 2008.

[103] M. Stevens, J. Pollak, S. Ralph, and M. Snorrason, "Video surveillance at night," in *Acquisition, Tracking, and Pointing XIX*, vol. 5810 of *Proceedings of SPIE*, pp. 128–136, 2005.

[104] Y. Zhao, H. Gong, L. Lin, and Y. Jia, "Spatio-temporal patches for night background modeling by subspace learning," in *Proceedings of the 19th International Conference on Pattern Recognition (ICPR '08)*, December 2008.

[105] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 162–182, 2007.

[106] S. Iwasawa, K. Ebiharai, J. Ohya, and S. Morishima, "Real-time estimation of human body posture from monocular thermal images," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 15–20, June 1997.

[107] B. Bhanu and J. Han, "Kinematic based human motion analysis in infrared sequences," in *Proceedings of the 6th IEEE Workshop on Applications of Computer Vision*, pp. 208–212, 2002.

[108] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 1, pp. 63–71, 2005.

[109] H. Nanda and L. Davis, "Probabilistic template based pedestrian detection in infrared videos," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, vol. 1, pp. 15–20, 2002.

[110] E. Goubet, J. Katz, and F. Porikli, "Pedestrian tracking using thermal infrared imaging," in *Infrared Technology and Applications XXXII*, vol. 6206 of *Proceedings of SPIE*, pp. 797–808, 2006.

[111] H. Zhao and S. S. Cheung, "Human segmentation by fusing visiblelight and thermal imaginary," in *Proceedings of the International Conference on Computer Vision (ICCV '09)*, 2009.

[112] P. Kumar, A. Mittal, and P. Kumar, "Fusion of thermal infrared and visible sprectrum video for robust surveillance," in *Proceedings of the 5th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP '06)*, vol. 4338 of *Lectures Notes in Computer Science*, pp. 528–539, Madurai, India, December 2006.

[113] J. Han and B. Bhanu, "Detecting moving humans using color and infrared video," in *Proceedings of the International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 228–233, 2003.

[114] L. Jiang, F. Tian, L. E. Shen et al., "Perceptual-based fusion of IR and visual images for human detection," in *Proceedings of the International Symposium on Intelligent Multimedia, Video and Speech Processing (ISIMP '04)*, pp. 514–517, October 2004.

[115] D. Rabinkin, R. Renomeron, A. Dahl, J. French, J. Flanagan, and M. Bianchi, "A DSP implementation of source location using microphone arrays," *The Journal of the Acoustical Society of America*, vol. 99, no. 4, pp. 2503–2529, 1996.

[116] P. Aarabi and S. Zaky, "Robust sound localization using multi-source audiovisual information fusion," *Information Fusion*, vol. 2, no. 3, pp. 209–223, 2001.

[117] B. Bhanu and X. Zou, "Moving humans detection based on multimodal sensor fusion," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2004.

[118] E. Menegatti, E. Mumolo, M. Nolich, and E. Pagello, "A surveillance system based on audio and video sensory agents cooperating with a mobile robot," in *Proceedings of the 8th International Conference on Intelligent Autonomous Systems (IAS '08)*, 2004.

[119] N. Megherbi, S. Ambellouis, O. Colôt, and F. Cabestaing, "Joint audio-video people tracking using belief theory," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS '05)*, pp. 135–140, September 2005.

[120] M. Cristani, M. Bicego, and V. Murino, "Audio-Video integration for background modelling," in *Proceedings of the 8th European Conference on Computer Vision (ECCV '04)*, vol. 3022 of *Lecture Notes in Computer Science*, pp. 202–213, Prague, Czech Republic, May 2004.