Advanced background modeling with RGB-D sensors through classifiers combination and inter-frame foreground prediction

Massimo Camplani · Carlos Roberto del Blanco · Luis Salgado · Fernando Jaureguizar · Narciso García

Abstract An innovative background modeling technique that is able to accurately segment foreground regions in RGB-D imagery (RGB plus depth) has been presented in this paper. The technique is based on a Bayesian framework that efficiently fuses different sources of information to segment the foreground. In particular, the final segmentation is obtained by considering a prediction of the foreground regions, carried out by a novel Bayesian Network with a depth-based dynamic model, and, by considering two independent depth and color-based mixture of Gaussians background models. The efficient Bayesian combination of all these data reduces the noise and uncertainties introduced by the color and depth features and the corresponding models. As a result, more compact segmentations, and refined foreground object silhouettes are obtained. Experimental results with different databases suggest that the proposed technique outperforms existing state-of-the-art algorithms.

Keywords Background modeling · Foreground prediction · Mixture of Gaussian · RGB-D cameras · Microsoft Kinect · Classifier combination

1 Introduction

Background modeling is one of the main tasks of video processing and analysis that aims at identifying a robust model of the static environment (the background), and consequently to detect the moving objects (the foreground) present in the scene. It is a very important task in many video-based applications such as: enhanced video conference systems, surveillance, advanced sport games monitoring, etc. In general, background modeling is used to process the data provided by static cameras in both indoor and outdoor scenarios, where the background model is iteratively built, and then any data deviation of this model is considered as a part of the foreground.

Different background modeling techniques (also called background subtraction or background/foreground segmentation) have been presented in literature as reported in the recent surveys presented in [7,12], which try to solve the different problems and challenges that strongly affect their performance. In particular, as presented in [12] the main challenges for background modeling algorithms are: *stopping foreground objects*, *multimodal background*, and *bootstrapping* (or initialization). *Color camouflage* is another challenge that occurs when the color features of the model are similar to the ones of the moving object, which usually cause fragmented foreground objects. Another important aspect that affects the performance of background subtraction algorithms is the *illumination change* (gradual or sudden): a modification of the static background could generate false foreground detections. *Shadow* cast by foreground objects in the

scene can also affect negatively the identification of the foreground regions. Finally, the problem of *moved background objects* occurs when background objects are moved from their original position, thus generating a new *empty* space that can be erroneously identified as foreground.

Recently, low-cost RGB-D cameras, such as Microsoft Kinect, have generated great interest in the world of computer vision since they guarantee real-time registered depth and color information. These devices have been rapidly employed in several computer vision applications such as sign recognition for human computer interfaces [27], home care activity monitoring [26], people re-identification in video surveillance [3], and people detection [28]. In these applications, especially in human-computer interfaces, the depth data relative to the moving objects (users) is segmented from the static scene, and then processed and analyzed. For this reason, efficient background modeling algorithms that use depth and color data are required to improve and broaden the possible applications for this kind of devices.

Depth information generated by RGB-D cameras is very efficient for background subtraction algorithms since it does not suffer the typical color imagery problems, as demonstrated in the recent review [17]. In particular, the effects of illumination changes, shadows, and color camouflage can be dramatically reduced with the use of depth data. On the contrary, depth data presents several problems that have to be carefully considered in the design of depth-based background modeling algorithms. Specifically, object silhouettes are heavily affected by the high level of noise at object boundaries, as shown in [10,11]. Depth measurements may be not available for all the pixels due to occlusions, multiple reflections, very distant points, or particular reflective surfaces, such as windows or television screens (see [11]). Moreover, as demonstrated in [20], a quadratic relationship between the noise variance and the measured depth exists. Finally, it should be considered that RGB-D cameras based on structured light scanner (i.e., Microsoft Kinect) are not suitable for outdoor environments, due to the range limitation and errors introduced by interferences with the sunlight.

As previously mentioned, several background modeling algorithms have been proposed in literature to address these challenging issues [12]. However, there is still a lack of research on methods that employ both color and depth data provided by RGB-D devices. This situation is exhibited in the recent review articles [7,12] in which very few works use depth data.

The algorithm presented in [15] combines color and depth data acquired by a stereo devices. The mixture of Gaussians (MoG) algorithm [29] is employed to model the background by a combination of four-dimensional Gaussian distributions per pixel. One component is the depth, and the other three ones are the color features (YUV space). In this study, depth and color features are considered independent. The original MoG algorithm is adapted so that if a reliable distribution match is found in the depth component; the color-based matching criterion is relaxed, resulting in the reduction of color camouflage errors. On the contrary, if the stereo matching algorithm is not reliable, the color-based matching criterion is set to be harder to reduce shadows or local illumination changes.

The MoG algorithm has been also used in [31], where depth and infrared data are combined for moving object detection. In this work, two independent background models are estimated, and the corresponding foreground regions are identified when the classification of these two models agree. Therefore, the errors due to one model failure could affect negatively the final classification.

A similar approach has been proposed in [23], where the color and depth data acquired by a Time-of-Flight (ToF) camera are combined for video segmentation. The Vibe algorithm [4] is employed to combine the two independent models and to obtain the foreground masks, which are combined with logical operations, and processed with morphological filters.

Depth-based background subtraction has been also used in the 3DTV content generation system presented in [14]. A depth-based model of the background scene is obtained through the MoG algorithm, and the detected foreground objects are removed from the applied pre-processing steps to improve the accuracy of background depth data. Moreover, foreground regions could be projected in virtual environments.

Few works have been developed for stereo-based or ToF technologies, and they do not consider the noise characteristics of the depth data provided by RGB-D cameras. Examples based on new RGB-D devices, such as [30] and [26], rely only on the depth data without considering a possible color and depth data integration.

A more efficient integration of depth and color has been proposed in the recent work presented in [9], where per-pixel statistical classifiers (based on depth and color data) are fused with a weighted average combiner. A mixture of Gaussian distribution is used to model the background pixels, and a uniform distribution is used for the foreground model.

In this paper, we present an innovative background modeling strategy that is able to obtain an accurate segmentation of foreground regions from data provided by RGB-D cameras in challenging indoor environment. In particular, the proposed approach allows to efficiently tackle strong illumination variations, interference due to the existence of multiple active RGB-D cameras, depth data noise and non measured depth data due to reflections and out-of-range problems of the RGB-D cameras, and situations of sudden people crowds. The proposed strategy is based on a Bayesian framework that efficiently fuses different sources of information to segment the foreground; the generation and the combination of these sources of information are precisely the main

contributions of the paper. The first source of information considered in the combination is a prediction of the foreground between consecutive images that relies on a novel adaptive block-based foreground modeling, which employs a depth dynamic model to robustly predict the temporal evolution of deformable foreground regions. Two depth-based and color-based independent per-pixel background models are also included in the combination. Each model is built starting from the MoG background modeling algorithm. The advantage of this approach is that the current contribution of the depth and color features and their corresponding models for the final foreground segmentation are solved by the Bayesian framework. In addition, the combination of the per-pixel background models and the region-based predicted foreground guarantees more compact and accurate segmentations. It is worth noting that the use of the foreground prediction is a very innovative point of the proposed work, since the state-of-the-art RGB-D based algorithms rely only on the background model. The proposed method has been tested in indoor environments due to the limitations of the RGB-D imagery in outdoor environments, as illustrated at the beginning of this section.

The rest of the paper is structured as follows: In Sect. 2, the proposed background modeling strategy is presented. Results are shown in Sect. 3. And lastly, the conclusions are drawn in Sect. 4.

2 Proposed Bayesian background modeling approach

The block diagram of the proposed background modeling approach is presented in Fig. 1. As it can be noticed, it is composed of three main blocks: the background modeling block (*BgMOD*), the foreground prediction block (*FgPRED*) and the Bayesian combiner block (*Combiner*). *BgMOD* is in charge of updating continuously the two independent models (based respectively on the depth data D and color features C), through the MoG algorithm. The second block, *FgPRED*, analyzes the previous foreground segmentation (Fg_{t-1} in Fig. 1) given by *Combiner*, and the current (D_t) and past (D_{t-1}) depth data to predict the position of the foreground regions (probabilities p_{fg} in Fig. 1). Finally, the *Combiner* block fuses the information provided by the other modules in a Bayesian framework to obtain an accurate final foreground detection. The likelihood probabilities (L_C and L_D) provided by the *BgMOD* block are combined with the foreground prediction p_{fg} estimated by *FgPRED* to calculate a per pixel posterior probability based on the two different features. This statistical information is finally combined with a weighted average that takes into account the color and depth-edges and the presence of non measured pixels. More details about these blocks are given in the following sections.

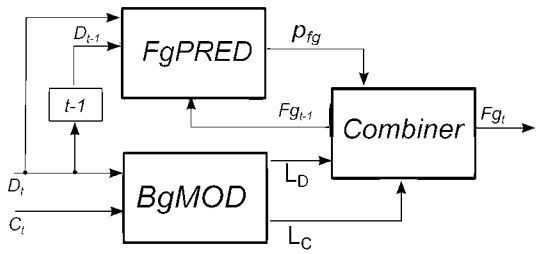


Fig. 1 Block Diagram of the proposed background modeling approach. D_t is the depth and C_t is the color data, past depth data D_{t-1} and past foreground detection Fg_{t-1}. p_{fg} represents foreground prediction probabilities, and L_C and L_D are the color-based and depth-based likelihood probabilities estimated by *BgMOD*

2.1 Depth and color background modeling: *BgMOD*

The main idea of the *BgMOD* block presented above is to consider the depth and color data as independent features, and consequently to use them to build two independent background models. These models are obtained by applying the MoG background modeling algorithm on depth data D and color data C. From now on, we will refer to the classifiers built on these models as MoG_C and MoG_D. The MoG algorithm is a very popular per-pixel background modeling technique (see the comprehensive review presented in [8] for more details), which allows to cope with multi-modal backgrounds, and to adapt the model to gradual changes. Each pixel is modeled as a mixture of Gaussian distribution and their parameters are learned, and iteratively updated through an online version of the Expectation Maximization algorithm.

In the MoG model, the probability to find a pixel at position s at time t of intensity $\mathbf{x}$ is defined as a mixture of Gaussians:

$$P\left(\mathbf{x}_{s,t}|\omega_{bg},\omega_{fg}\right)=\sum_{i=1}^{K} v_{i,t}\cdot\mathcal{N}\left(\mathbf{x}_{s,t},\mu_{i,t},\Sigma_{i,t}\right) \quad (1)$$

where K is the number of Gaussians used in the model, $v_{i,t}$ is the weight associated to the ith Gaussian $\mathcal{N}$ at the time t with mean $\mu_{i,t}$ and covariance matrix $\Sigma_{i,t}$. It is worth noting that the statistics of both models, background and foreground, are described with the same mixture, which is indicated in Eq. 1 by inserting the two classes symbols, ω_{bg} (for the background) and ω_{fg} (for the foreground), in the left term of the equation.

As far as the distribution dimensionality is concerned, the depth-based distributions have a single dimension and the color ones have three components. As widely proposed in the literature (see the reviews [7,8]), the color components are considered independent, and therefore Σ is a diagonal matrix containing the variances of the three color components. This assumption allows to reduce the complexity of

the MoG algorithm. Regarding the weight $v_{i,t}$, it measures the accuracy with which the corresponding Gaussian models the corresponding pixel.

In the first step of the MoG algorithm it is checked if the pixel belongs to the background model or to the foreground distribution. The background model is formed by those distributions of the mixture that are characterized by a high ratio (r) between the weight and the variance. In the case of a univariate distribution r, the ratio for the ith Gaussian is estimated as $r_i = v_i/\sigma_i$. A high value for r_i means that the corresponding distribution is characterized by a low variability (a low value of σ) and a high weight v_i. The distributions are ordered by considering the factor r and the background is composed by the first B distributions that exceed the threshold T:

$$B = \arg\min_b \left(\sum_{i=1}^{b} v_{i,t} \geq T \right) \tag{2}$$

T indicates the minimum portion of data that should be accounted for by the background. For small values of T, a background modeled by few distributions is obtained (at least a unimodal Gaussian distribution). If T is high, a multimodal background model is obtained, that can include more than one distribution in it. The Mahalanobis distance is used to check if the pixel belongs to one of the distributions:

$$\sqrt{\left(\mathbf{x}_{s,t+1} - \mu_{i,t}\right)^T \Sigma_{i,t}^{-1} \left(\mathbf{x}_{s,t+1} - \mu_{i,t}\right)} < \lambda, \tag{3}$$

where λ is usually set to 2.5 (see for example [8]). If Eq. 3 is satisfied for one of the background distributions, the pixel will be classified as a background pixel, otherwise it is classified as a foreground pixel.

As mentioned in the introduction, the depth measurement variation follows a quadratic relationship with the depth. Therefore, depth measurements corresponding to regions located far from the camera are characterized by a higher value of σ than those regions situated closer. This fact could introduce a bias in the estimation of the ranking parameter r. For this reason, we normalize its value with σ_{noise}, as presented in [11], that is selected according to the quadratic law between distance and noise dispersion presented in [20].

The parameters of the background distributions are updated as proposed in [29]. In particular, in the case of a single dimension feature space (i.e. depth):

$$v_{i,t+1} = v_{i,t}(1-\alpha) + \alpha \tag{4}$$

$$\rho = \alpha \cdot \mathcal{N}\left(x_{s,t}, \mu_{i,t}, \sigma_{i,t}\right) \tag{5}$$

$$\mu_{i,t+1} = \mu_{i,t}(1-\rho) + \rho x_{s,t+1} \tag{6}$$

$$\sigma_{i,t+1}^2 = \sigma_{i,t}^2(1-\rho) + \rho\left(x_{s,t+1} - \mu_{i,t+1}\right)^2 \tag{7}$$

where α is the *learning rate*, which influences the speed of background modifications due to new objects in the scene or gradual changes (i.e. changes of the illumination conditions). It indicates the impact that the last samples have on the updated distributions' parameters. On the other hand, for the unmatched Gaussians all the parameters remain unchanged except their weight:

$$v_{i,t+1} = v_{i,t}(1-\alpha). \tag{8}$$

In the case that Eq. 3 has not been satisfied by all the distributions, the one with the lowest ratio r is substituted by a new one with low weight, high variance, and a mean equal to the current value. It is worth noting that in each iteration the weights are normalized such that their sum is equal to one.

Instead of using a fixed value for α, as described in the original version of the MoG algorithm [29], we use a variable learning rate as proposed in [19]. Its value is set to $1/N_{frame}$, where N_{frame} is the number of processed frames, until a preset minimum value of α is reached. In this way, the first frames have a high impact on the parameters evolution, thus avoiding the incorporation of moving objects present in the beginning of the sequence to the background.

Moreover, we apply a control at frame level in order to re-initialize the MoG_C module in case of sudden change of luminosity. As proposed in [32], we check the fraction of pixels that have been classified as belonging to the foreground. If this fraction is greater than a threshold (i.e. $\approx$0.6) an illumination change is detected and the MoG_C module has to be re-initialized. It is worth noting, that low cost RGB-D cameras do not allow to control acquisition settings such as aperture time, shutter time, white balance, etc. Due to the automatic modification of these parameters, the color data is affected by sudden changes of the illumination condition (also in very controlled environment) that affect large portion of the image.

The MoG algorithm guarantees a reliable online estimation of both background and foreground distributions. By considering the estimated distribution parameters, it is possible to compute the likelihood probability that a pixel belongs to these two models. The estimated likelihood factors for both classes (ω_{fg} and ω_{bg}) and for both features (D and C) are then combined to improve the quality of the foreground detection.

2.2 Prediction of foreground regions: *FgPRED*

A probability map p_{fg} containing the prediction of the foreground regions at the current time step is estimated using Fg_{t-1} (the foreground segmentation of the previous time step), a dynamic model of the foreground, and another model related to the depth-based appearance of the foreground. The proposed background subtraction strategy uses this probability map as prior knowledge to improve the probabilistic modeling of the foreground.

The foreground segmentation Fg_{t-1} is divided into squared sub-regions of fixed size. The center coordinates of

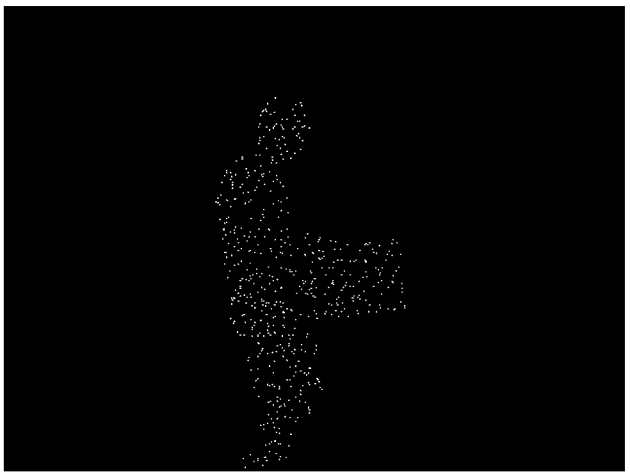


Fig. 2 Division of Fg_{t-1} into sub-regions. Each point represents the center of a square region

the square sub-regions are randomly sampled from the set of pixels belonging to Fg_{t-1}. The number of sub-regions is proportional to the area of the segmented foreground, so that the whole set of square sub-regions completely cover with a certain overlapping the foreground. This division allows to predict more accurately the foreground regions, especially for those foreground regions that contain deformable objects. Figure 2 shows the center of the square sub-regions obtained after the sampling-based division.

The position of each foreground sub-region is predicted on D_t (the current depth image) using the following Bayesian Network model described. A foreground sub-region is parameterized by the variables $\{u, v, d\}$, where u and v are the coordinates of the sub-region center, and d is the shift in depth of the sub-region. This variable, representing the shift in depth, is required since the relative depth of a sub-region can vary between consecutive time steps because of its motion. System observations, represented by the variable h, are based on an appearance model that uses a depth histogram to characterize a sub-region. Given h, the posterior pdf on $\{u, v, d\}$ is proportional to

$$p(u, v, d|h) \propto p(h|u, v, d)p(u, v, d), \tag{9}$$

where $p(u, v, d)$ is the prior probability on $\{u, v, d\}$ in the current time step, and $p(h|x, y, d)$ is the likelihood that a candidate sub-region forms part of the foreground using the information of the depth histogram.

The prior probability $p(u, v, d)$ follows a zero-order model that encodes the idea that the temporal evolution of the position of a sub-region should be smooth and bounded. This assumption is fulfilled for the target application of pedestrian segmentation and the used camera settings. For other specific camera settings that do not verify the previous assumption, it is recommended to utilize a higher frame rate, or place the depth camera in such a way that pedestrians are not so close to it. The prior probability is mathematically expressed by the following multi-variate Normal distribution

$$p(u, v, d) = \mathcal{N}(u, v, d; \mu_u, \mu_v, \mu_d, \Sigma_{u,v,d}). \tag{10}$$

The means are $\{\mu_u = u', \mu_v = v', \mu_d = 0\}$, where u' and v' are the center coordinates of the initial squared sub-region resulting from the foreground subdivision. Note that $\mu_d = 0$ because d is a shift measure, i.e. a relative quantity of the depth, not an absolute quantity. The covariance matrix $\Sigma_{u,v,d}$ is defined by its diagonal $[\sigma_u^2, \sigma_v^2, \sigma_d^2]$ since the involved variables are considered independent.

The likelihood term $p(h|u, v, d)$, is based on the similarity between the depth histogram of the candidate sub-region in the current time step and the evolution of the depth histogram of the initial foreground sub-region in the previous time step. The depth histogram of a foreground sub-region evolves in time since its relative depth changes due to its motion, either because it is approaching the camera sensor, or because it is moving away from it. The depth evolution between consecutive time steps, given by d, is used to estimate h' that is a prediction of the depth histogram for the current time step. This predicted histogram is obtained using a linear interpolation process based on the assumption that the sub-region locally experiences the same shift in depth d. The similarity between the predicted depth histogram and the corresponding one to the candidate sub-region is computed using the Bhattacharyya distance

$$b_d = \sqrt{1 - b_c}, \tag{11}$$

where b_c is the Bhattacharyya coefficient given by

$$b_d = \sum_i \sqrt{h(i)h'(i)}. \tag{12}$$

The likelihood is finally computed as

$$p(h|x, y, d) = N(b_d; 0, \sigma_h^2), \tag{13}$$

where the variance σ_h^2 expresses the expected uncertainty of the observation model.

The posterior pdf $p(u, v, d|h)$ cannot be directly computed because of the lack of an analytical expression due to the non-linearities and multi-modalities of the dynamic and observation models [2]. For this reason, an approximate inference method called Metropolis-Hastings algorithm [6] is used to obtain a discrete approximation of the posterior pdf as

$$p(u, v, d|h) = \sum_i \delta(u - u^i, v - v^i, d - d^i), \tag{14}$$

where $\{u^i, v^i, d^i\}$ are the samples from the posterior pdf. The Metropolis-Hasting algorithm generates these samples by means of a Markov chain that tries to simulate the target pdf. For this purpose, the following proposal distribution is used to draw the samples

$$\{u^i, v^i, d^i\} \sim q(u, v, d) = \mathcal{N}(u, v, d; u', v', 0, \Sigma'_{u,v,d}), \tag{15}$$

where $\Sigma'_{u,v,d}$ is a diagonal matrix defined by the vector $[\sigma'^2_u, \sigma'^2_v, \sigma'^2_d]$. Notice that this proposal distribution is similar to the aforementioned prior distribution, but the values of their covariance matrices are very different because of their distinct purposes.

Once that the posterior pdf of each foreground sub-region has been approximated by a set of samples, a unique set of samples is created by means of the union of all the samples, which represents a discrete approximation of posterior pdf of the whole foreground regions

$$p_{fg}(u, v, d|h) = \bigcup_j p_j(u, v, d|h), \tag{16}$$

where j is an index used to enumerate the posterior pdf of each foreground sub-region. Figure 3 shows the sampled-based approximation of $p_{fg}(u, v, d|h)$ marginalized over the variable d. Note the high concentration of samples on the foreground region, in this case a person.

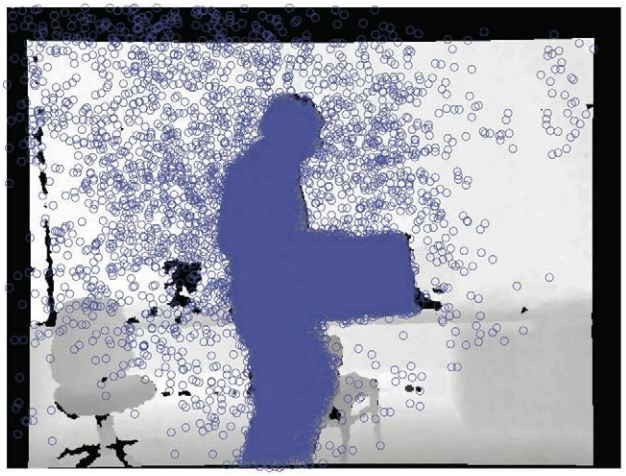


Fig. 3 Sampled-based approximation of $p_{fg}(u, v, d|h)$ marginalized over the variable d

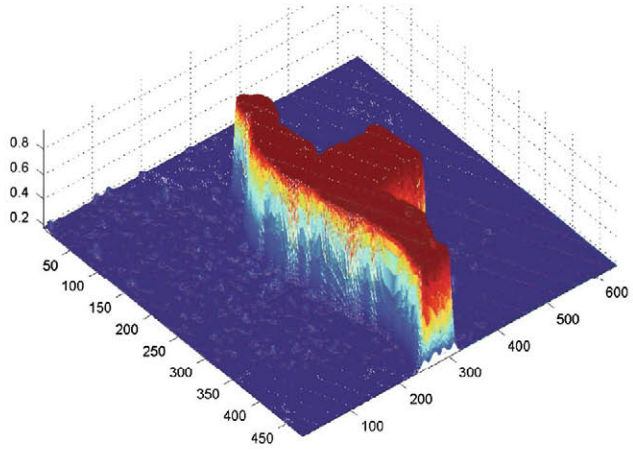


Fig. 4 Probability map of foreground regions after the smoothing of $p_{fg}(u, v, d|h)$

Lastly, a Gaussian-based kernel smoothing technique [5] is applied to $p_{fg}(u, v, d|h)$ to obtain a dense probability map that contains the prediction of the foreground regions at the current time step. Figure 4 shows the resulting probability map of foreground regions after the smoothing of $p_{fg}(u, v, d|h)$.

2.3 Bayesian combiner

The problem of background modeling and foreground extraction can be viewed as a two-class problem in which for each pixel the label of the foreground class (ω_{fg}) or of the background one (ω_{bg}) has to be assigned. As previously mentioned, we propose to build two different background models based on the depth and color features. The final classification (pixel belonging to foreground or background) is improved by an efficient combination of the two *weak* classifiers (based on the background models) responses and the probabilities provided by the *FgPRED* block that analyzes and predicts the evolution of the foreground regions.

In many complex pattern recognition tasks, the combination of different classifiers has been employed with very positive results. In particular, it has been demonstrated that complex problems can be efficiently tackled by combining more simple classifiers that work on a *local* level instead of training and building a unique sophisticated classifier that processes all the features. This research area has been extensively treated in the literature; for a complete review see [22] and [13, Ch.9].

The per-pixel classifiers MoG_D and MoG_C estimate the likelihood (respectively L_D and L_C in Fig. 1) that a pixel belongs to the class ω_{fg} or ω_{bg}. This information can be combined with the data provided by *FgPRED* in order to estimate the posterior probability that the measured data belongs of one of the two classes. In particular, we can calculate the class prior probabilities such that $P(\omega_{fg}) = p_{fg}$ and $P(\omega_{bg}) = 1 - p_{fg}$ and consequently obtain, for each classifier, through the Bayes theorem, the posterior probability.

For each classifier, it is possible to calculate the vector $d = [d_{bg}, d_{fg}]$ that contains the estimated posterior probabilities for the two classes. Considering the pixel s at position (x, y), and the corresponding measured data $\mathbf{x}_s = [D, C]$, the decision profile DP [22] for $\mathbf{x}_s$ is:

$$DP(\mathbf{x}_s) = \begin{bmatrix} d_{C,bg} & d_{C,fg} \\ d_{D,bg} & d_{D,fg} \end{bmatrix} \tag{17}$$

where each row represents the vector d mentioned above, and each column represents the overall support $M(\mathbf{x}_s)$ of both classifiers to one of the classes. By analyzing $DP(\mathbf{x}_s)$, it is possible to extract the overall support for all the classes of the

problem at hand, and consequently assign the class-label to $\mathbf{x}_s$ corresponding to the class with the greatest value of $M_j(\mathbf{x}_s)$. In our case we select a *class conscious* combiner to fuse the data contained in DP matrix. These combiners are not trainable, and therefore it is not possible to estimate $M_j(\mathbf{x}_s)$ by using arithmetic operations; hence, their computational and memory requirements are generally low.

Typical choices of the combination are average, median, maximum, etc. (see [22, Ch.5] for more details), and they have been employed in many pattern recognition problems and also in computer vision. In particular, in the area of background modeling, an average combiner has been used in [21], where 13 classifiers are trained on 13 different visual cues. Simple combination functions (i.e. average) assign to all the features the same support to the final classification. In our case, we use a weighted average in order to extract different information from the different feature sets and to adapt efficiently the *contribution* of each classifier to the final classification. $M_j(\mathbf{x}_s)$ is estimated as:

$$M_j(\mathbf{x}_s) = W_C \cdot d_{C,j} + W_D \cdot d_{D,j} \tag{18}$$

The weights are chosen as a function of the input $\mathbf{x}_s$ to increase the support of the most reliable classifier. In the following paragraph a brief description on weights selection is presented.

First of all, for all those pixels for which the depth measurements is not available (called *nmd* from now on), the depth-based classifier weight (W_D) is set to 0, and the color-based classifier weight (W_C) is set to 1. In fact, when depth data is not available in the current frame or in the background model, only the color feature is considered for the final pixel classification.

For the pixels that do not belong to the *nmd* set, we assign the weights as a function of the depth-image edges as proposed by [9]. This is due to the fact that depth data guarantee generally compact detection of moving object regions except that for the very noisy depth values at object boundaries. To reduce this effect, we increase the influence of the color based classifiers in these regions.

For all those pixels that have valid depth data, we assign the weights as a function of the depth-image edges following the strategy proposed in [9]. In this way, it is possible to limit the effect of noisy depth values at object boundaries by using the color information in these zones. On the contrary, the depth information is more reliable in the regions far from depth-edges, since it guarantees compact foreground and it is resilient to shadows and illumination changes.

For depth and color data, the edge-closeness probability P^e is estimated for each pixel; it is calculated as a function of the distance between the pixel and the closest edge (previously detected with a conventional edge detector). This distance is weighted with a Gaussian function with standard deviation σ_{edge}. With the depth-edge-closeness probability (P^e_D) and the color edge-closeness probability (P^e_C) we estimate the global edge-closeness probability such that $P^e_G = P^e_C \cdot P^e_D$. The obtained P^e_G has a high value in those regions in which depth-based edges corresponds to color based edges. As mentioned in the introduction, depth-based edges are characterized by the presence of noise, hence color features are in general more reliable in these regions. For this reason, a higher weight has to be assigned to MoG_C in order to increase the final detection accuracy. It should be noticed that the product between the two edge-closeness probability functions limits the impact of MoG_C where there is not an edge in the color domain. The weights are assigned as $W_C(\mathbf{x}_s) = P^e_G(x_s)$ and $W_D(\mathbf{x}_s) = 1 - W_C(\mathbf{x}_s)$. The weights values are bounded to a minimum and a maximum value ($W_{\min}$ and $W_{\max}$) to ensure the contribution of both classifiers in the final classification. These values have been set in our implementation respectively to 0.1 and 0.9.

Finally, we propose a temporal consistency score tc for the depth-based background model region affected by the presence of *nmd* pixels. Let us consider an area corresponding to a high reflective region (i.e. a TV screen) in which several *nmd* pixels are present and, hence, there is no depth-based model available. If a foreground object temporally moves in front of this area, the depth-based model could be initialized with these depth values, that obviously do not correspond to the real depth of the tv screen in the background. Clearly, any other foreground object moving in front of the screen with similar depth values will be wrongly classified as background by the MoG_D. In order to reduce the effects of this problem to the final classification, we multiply the element $d_{D,bg}$ of the decision profile by the factor tc calculated as

$$tc = \frac{1}{1 + e^{\beta \# Hit}}, \tag{19}$$

where $\#Hit$ represents a counter that is incremented when a valid depth measurement is obtained, and is decremented when a *nmd* pixel is registered. Let us consider the parameter Hit_{max} as the value for the counter $\#Hit$ for which the tc value is equal to one (depth-based background model is completely reliable). The parameter $\#Hit$ indicates how many consecutive valid depth measurements are required to consider the depth-based model completely reliable ($tc = 1$). Once this parameter is selected, the value of the parameter β can be easily selected. Thanks to the temporal consistency score, it is possible to give a different weight to the background model relative to those regions that are strongly affected by the presence of *nmd* pixels.

3 Results

In this section, we present the results obtained with the proposed strategy and other state-of-the-art background model-

ing techniques based on depth and color features. For this purpose, we have used three different databases of indoor sequences. The first one [28] is a public database that is composed of three different indoor sequences acquired by an array of three RGB-D sensors (Kinect devices) in a university building hall. These sequences are particularly challenging because of the presence of crowded situations, the strong variations in illumination, and the clutter introduced by the interference between multiple RGB-D devices. The last problem arises because Kinect is an active sensor that emits infrared structured light, which can be wrongly sensed by other devices in the array. This problem strongly affects the sequences contained in this database, since the three Kinect devices are set in such a way that their fields of view overlap. Another source of problems is the lack of depth information in large areas of the acquired images due to reflections or objects that are positioned out of range of the sensor. For each sequence of this database, we have generated a hand-labeled ground truth containing 80 frames, which span over 400 frames of the sequence. The other database is the one proposed in [9] that includes scenes with a single person moving. The scenes contained in this database are less complex than the ones in [28]. For this reason, we included only one sequence (called *genSeq*) of this dataset in our tests. The last database [1] is composed by four sequences acquired by two Kinect cameras, which are placed on the ceiling looking forward with a significant pitch angle. These aspects differentiate these sequences from the others where the cameras are positioned parallel to the floor at a height between 1 and 1.5 meters. Color, depth, and ground truth information are available for each sequence, which usually contains over 40 different people. In our experiments we use two sequences of this dataset.

As objective measures to test the algorithm, we have used six well-known performance indexes: false positive rate (FP), false negative rate (FN), total error (TE), similarity measure S, and overall rankings RM and RC. FP is calculated as the fraction of the background pixels that are marked as foreground. FN is computed as the fraction of foreground pixels that are marked as background. TE is the total number of misclassified pixels, normalized with respect to the image size. The metric S is defined as [24]

$$S(A, B) = \frac{A \cap B}{A \cup B} \tag{20}$$

where A is the detected region and B is the ground truth region. This non-linear measure is a combination of FP and FN indexes. Values close to 1 indicate that A and B regions are very similar, and values close to 0 just the opposite. The last metrics, RM and RC, rank the overall accuracy of the tested methods [16]. Defining $\text{rank}_i(m, \text{sq})$ as the rank of the ith method for the performance metric m in the sequence sq, the average ranking of the method i in the sequence sq is expressed as

$$\text{RM}_i = \frac{1}{N_m} \sum_m \text{rank}_i(m, \text{sq}) \tag{21}$$

where N_m is the number of metrics used. The overall ranking RC_i for the ith method is then computed taking into account all the sequences

$$\text{RC}_i = \frac{1}{N_{\text{sq}}} \sum_{\text{sq}} \text{RM}_i, \tag{22}$$

where N_{sq} is the number of sequences. This rank indicates the global performance of one method with respect to the others.

The proposed algorithm (called $MoG\text{-}PRE$ for future reference) has been compared with five different algorithms of the state of the art: MOG_{Bin} [31], which computes a binary combination of foreground masks obtained by two independent modules using MoG; $Vibe_{Bin}$ [23], which is based on a binary combination of foreground masks obtained by means of the ViBe algorithm [4]; $PBAS$ [18], which first models the background using the recent history of pixel values, and then computes the foreground using a decision threshold calculated dynamically for each pixel; $SOBS$ [25], which adopts a neural-network based approach to detect foreground objects without making any assumption about the pixel distribution; and CL_W [9], which uses a probabilistic classifier to fuse a set of foreground masks, which are computed by a mixture of Gaussian approach that uses color and depth information. Regarding the $PBAS$ and $SOBS$ algorithms, we have to state that they have been extended to use RGB-D imagery, since originally they only employed color imagery. Finally, we have also used three additional baseline algorithms for the comparison: $MOG_{RGB\text{-}D}$ [15], which is based on an RGB-D mixture of Gaussian model; and MoG_D and MoG_C, which are the mixture of Gaussian modules employed in our system to build the color and depth-based model.

An example of the *genSeq* sequence presented in [9] is shown in Fig. 5. As it can be noticed, the depth data are affected by different perturbations: *nmd* (pixels marked in red), noisy object boundaries, and spatial noise affecting

(a) (b)

Fig. 5 *genSeq* sequence: color (**a**) and depth data (**b**)

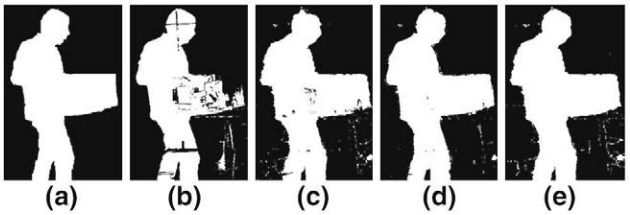


Fig. 6 Foreground detection mask for *genSeq* sequence: **a** ground truth,**b** MoG_C, **c** MoG_D, **d** $MoG\text{-}PRE$, **e** $MoG_{RGB\text{-}D}$

homogeneous depth regions. The corresponding foreground masks are reported in Fig. 6. The foreground silhouette obtained with the proposed strategy (Fig. 6d) is very compact and with well-defined boundaries. Moreover, the false detections due to the depth noise are severely reduced. It is worth noting that the results obtained with the single classifiers do not show the same accuracy. The MoG_C (Fig. 6b) is affected by the color camouflage problem, thus rendering a fragmented foreground mask with several holes. On the contrary, the foreground mask obtained with MoG_D (Fig. 6c) is more compact, although the detected object boundaries are affected by the depth noise. This example shows clearly how the proposed strategy allows to efficiently combine depth and color information, making use also of the foreground prediction strategy. The results obtained by $MoG_{RGB\text{-}D}$ are reported in (Fig. 6e). In this case, the obtained foreground mask boundaries are not well-defined, leading to more false detections.

The quantitative results using the previously introduced metrics are reported in Table 1. The proposed strategy achieves the best ranking score RM. It also guarantees the lowest level of TE and the highest value for S. Moreover, it improves both FP and FN metrics with regard to the values independently obtained by MoG_C and MoG_D, which are used in our system. As far as the other state-of-the-art algorithms is concerned, only CL_W has comparable results with our approach, although slightly inferior. On the other hand,

Table 1 Detection accuracy obtained by analyzing the *genSeq* sequence

	TE	FN	FP	S	RM
MoG-PRE	0.90	1.26	0.85	0.87	1.75
MoG_C	2.13	8.41	1.35	0.74	6.75
MoG_D	1.61	3.70	1.35	0.81	4.50
MoG_{Bin}	2.03	17.01	0.16	0.74	6.25
$MoG_{RGB\text{-}D}$	1.93	0.63	2.09	0.79	5.25
$ViBe_{bin}$	12.39	0.65	13.85	0.44	7.25
CL_W	1.13	2.26	0.99	0.85	3.25
SOB	1.91	1.34	1.98	0.80	4.75
$PBAS$	1.93	2.10	1.91	0.80	5.25

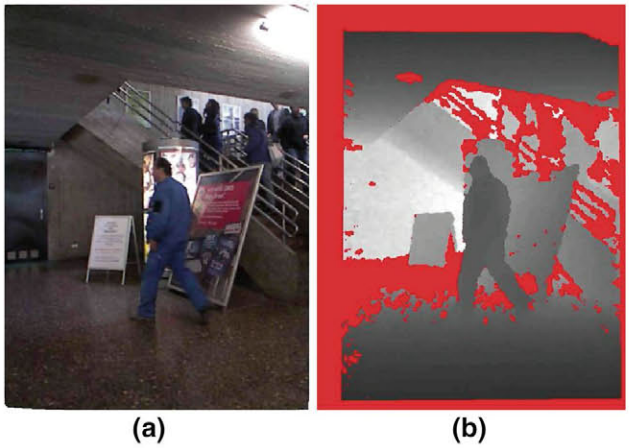


Fig. 7 Lobby1 sequence: color (**a**) and depth data (**b**)

the simple binary combination of the foreground masks computed by MoG_{Bin} and $ViBe_{Bin}$ does not guarantee accurate results, leading to unbalanced levels of FN and FP. In fact, the final decision is completely compromised when one of the classifier fails.

In Fig. 7, depth and color data of the first sequence (Lobby1) of the database presented in [28] are shown. As previously mentioned, the sequences in this database show large image areas for which the depth data is not available, due to reflective surfaces and out of range objects. The obtained foreground masks are reported in Fig. 8. The proposed strategy $MoG\text{-}PRE$ (Fig. 8d) improves the detections obtained by the two weak classifiers independently. In particular, the MoG_C (Fig. 8b) is affected by cast shadows on the floor and local changes of illumination (i.e. lighting advertising panel behind the man). On the contrary, the foreground mask obtained by MoG_D (Fig. 8c) is more compact, but it does not accurately segment distant moving objects (peoples on the stairs), and closer moving-object parts (for example the man's leg). As it can be noticed, the proposed strategy reduces the effect of the above mentioned problems. Also, the silhouettes obtained by $MoG_{RGB\text{-}D}$ are very compact, although the foreground detection is affected by a higher level of false positives.

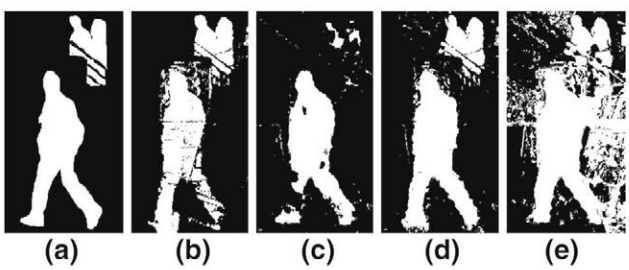


Fig. 8 Foreground detection mask for Lobby1 sequence: **a** ground truth,**b** MoG_C, **c** MoG_D, **d** $MoG\text{-}PRE$, **e** $MoG_{RGB\text{-}D}$

Table 2 Detection accuracy obtained by analyzing the Lobby1 sequence

	TE	FN	FP	S	RM
MoG-PRE	5.65	19.43	3.67	0.48	2.25
MoG_C	11.43	55.91	5.04	0.28	6.50
MoG_D	5.42	31.41	1.69	0.41	2.25
MoG_{Bin}	12.00	66.73	4.13	0.18	7.50
$MoG_{RGB\text{-}D}$	25.26	10.61	27.36	0.26	6.50
$ViBe_{bin}$	14.99	45.87	10.55	0.23	7.50
CL_W	7.72	15.45	6.61	0.45	3.50
SOB	8.75	44.82	3.57	0.33	3.75
$PBAS$	10.21	54.34	3.87	0.28	5.25

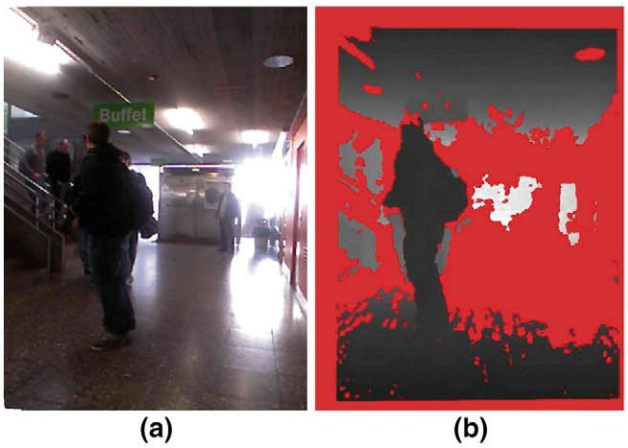


Fig. 9 Lobby2 sequence: color (**a**) and depth data (**b**)

The performance metrics for the Lobby1 sequence are reported in Table 2. In this case, the proposed algorithm $MoG\text{-}PRE$ shares the best ranking score RM with the MoG_D method, but $MoG\text{-}PRE$ guarantees an efficient combination of the color and depth data, leading to a better tradeoff performance with respect to the two independent classifiers. In fact, it reduces dramatically the value of FN, and at the same time guarantees a low value for FP. Moreover, it leads to the highest value of S. The CL_W method offers also comparable results to $MoG\text{-}PRE$, but slightly worse. Finally, the other state-of-the-art techniques do not provide comparable performance to the proposed approach.

Figure 9 shows the depth and color data of the second sequence (Lobby2) of the database proposed in [28]. An example of the obtained foreground masks is reported in Fig. 10. In this case, due to the particular lighting conditions and the presence of a very large region with *nmd* pixels, the two weak classifiers (see Fig.10b and c) present fragmented foreground masks. However, as it can be noticed in Fig.10d, the proposed strategy allows to dramatically improve the final foreground detection. This example shows the positive influence of two important aspects of the proposed strategy:

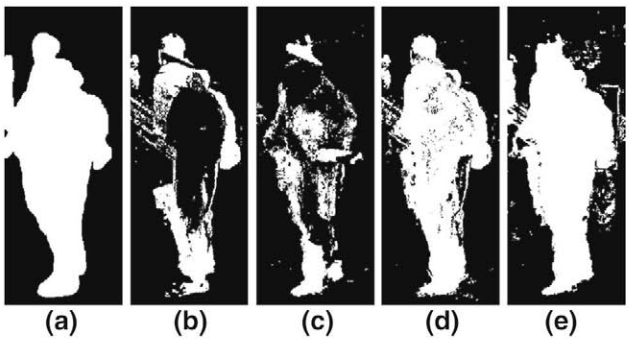


Fig. 10 Foreground detection mask for Lobby2 sequence: **a** ground truth, **b** MoG_C, **c** MoG_D, **d** $MoG\text{-}PRE$, **e** $MoG_{RGB\text{-}D}$

Table 3 Detection accuracy obtained by analyzing the Lobby2 sequence

	TE	FN	FP	S	RM
MoG-PRE	7.41	29.10	4.60	0.48	4.00
MoG_C	11.38	57.30	5.44	0.30	7.75
MoG_D	7.51	60.00	0.71	0.30	5.25
MoG_{Bin}	8.01	67.89	0.26	0.27	6.25
$MoG_{RGB\text{-}D}$	8.22	17.86	6.97	0.47	5.25
$ViBe_{bin}$	7.54	56.75	1.17	0.38	4.50
CL_W	5.21	24.32	2.74	0.54	2.00
$SOBS$	6.00	31.01	2.76	0.51	3.25
$PBAS$	9.13	57.43	2.88	0.34	6.75

the $FgPRED$ module and the weight selection strategy proposed in Eq. 19. In fact, both of them allow to increment the overall support M for the class ω_{fg}. Comparable results are obtained with $MoG_{RGB\text{-}D}$, where a compact foreground mask is obtained, although the object boundaries are not well-defined (i.e. man's head), and a higher level of false positives is obtained.

The comparison of all the methods is reported in Table 3. The proposed approach achieves the third best ranking score RM, after CL_W and $SOBS$, and it is very close to the second one. Similar results are obtained for the S and TE metrics. This fact can be explained by the lower level of false positives of the CL_W and $SOBS$ methods for this specific sequence.

An example of the last sequence (Lobby3) of the database proposed in [28] is shown in Fig. 11. The foreground mask obtained by the proposed method is reported in Fig. 12d, which has higher accuracy than that of the two independent classifiers (Fig. 12b, c), leading to more compact and defined objects silhouette. On the other hand, the method $MoG_{RGB\text{-}D}$ is strongly affected by illumination changes.

The results obtained with the different algorithms are presented in Table 4. The proposed method shares the second best ranking score RM with CL_W method. The first ranking score is achieved by MoG_D, which is slightly better in

Fig. 11 Lobby3 sequence: color (**a**) and depth data (**b**)

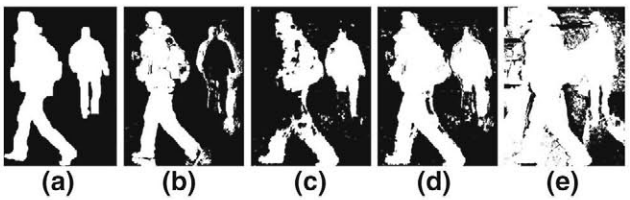


Fig. 12 Foreground detection mask for Lobby3 sequence:**a** ground truth, **b** MoG_C, **c** MoG_D, **d** $MoG\text{-}PRE$, **e** $MoG_{RGB\text{-}D}$

Table 4 Detection accuracy obtained by analyzing the Lobby3 sequence

	TE	FN	FP	S	RM
MoG-PRE	5.42	21.59	3.95	0.33	2.75
MoG_C	9.19	54.74	5.06	0.21	5.00
MoG_D	3.65	31.70	1.11	0.34	2.00
MoG_{Bin}	11.50	58.24	7.27	0.13	7.00
$MoG_{RGB\text{-}D}$	48.04	9.21	51.56	0.13	6.50
$ViBe_{bin}$	25.24	41.31	23.79	0.12	7.50
CL_W	7.06	18.06	6.06	0.36	2.75
$SOBS$	13.42	46.10	10.46	0.19	6.25
$PBAS$	8.44	66.46	3.18	0.17	5.25

terms of TE and S. This is mainly due to the particular challenging illumination conditions of this sequence. In fact, the sequence presents frequent changes of illumination due to the automatic white balance adjustment of the camera. This factor affects negatively the pixel classification stage (increasing the number of FP) in the large regions without depth information. The proposed algorithm also reduces the number of FN with respect to the two independent classifiers, while at the same time keeping a low value of FP.

Figure 13 shows a pair of color and depth images from the Cam1 sequence belonging to the dataset proposed by [1]. Similarly to the other sequences, there are large areas for which the depth data is not available due to the presence of reflective surfaces and out of range objects. Figure14 shows a detail of the foreground masks obtained with MoG_C, MoG_D, $MoG\text{-}PRE$, CL_W, $SOBS$, and $PBAS$, along with the ground truth. The proposed algorithm, $MoG\text{-}PRE$, achieves the best results, obtaining a compact segmentation with few false positives. The algorithm CL_W also achieves a compact segmentation but with significantly more false positives.

Fig. 13 Cam1 sequence: color (**a**) and depth data (**b**)

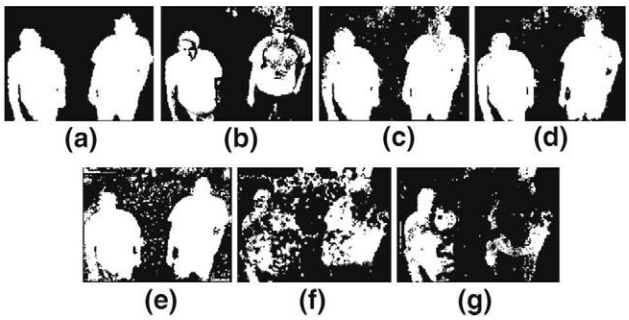


Fig. 14 Foreground detection masks for Cam1 sequence: **a** ground truth, **b** MoG_C, **c** MoG_D, **d** $MoG\text{-}PRE$, **e** CL_W, **f** $SOBS$, **g** $PBAS$

Table 5 Detection accuracy obtained by analyzing the Cam1 sequence

	TE	FN	FP	S	RM
MoG-PRE	6.25	13.44	5.33	0.61	1.75
MoG_C	10.42	46.14	5.57	0.40	5.00
MoG_D	7.30	17.37	5.98	0.55	3.50
MoG_{Bin}	8.36	55.16	1.97	0.39	5.00
$MoG_{RGB\text{-}D}$	32.91	0.89	37.36	0.33	6.75
$ViBe_{bin}$	13.76	48.59	8.97	0.30	7.75
CL_W	6.50	9.50	5.98	0.60	2.75
SOB	10.49	25.59	8.42	0.47	5.25
$PBAS$	14.74	43.53	10.78	0.34	7.25

The comparison of all the approaches using Cam1 and Cam2 sequences of the dataset proposed by [1] are shown in Tables 5, 6, respectively. The proposed approach achieves the best ranking score RM for both sequences. It also outperforms the other methods by considering TE and S metrics, proving that it guarantees the best tradeoff between the false positives and false negatives scores.

Table 6 Detection accuracy obtained by analyzing the Cam2 sequence

	TE	FN	FP	S	RM
MoG-PRE	6.87	8.90	6.55	0.63	2.00
MoG_C	11.30	42.27	6.11	0.42	4.75
MoG_D	7.97	12.01	7.35	0.59	3.75
MoG_{Bin}	9.30	48.58	2.75	0.44	4.50
$MoG_{RGB\text{-}D}$	23.35	1.28	26.84	0.42	6.25
$ViBe_{bin}$	13.55	38.99	9.27	0.39	7.00
CL_W	6.95	6.11	6.99	0.62	2.50
SOB	13.04	41.27	8.31	0.40	6.25
$PBAS$	13.96	45.18	8.73	0.36	8.00

Table 7 Algorithm ranking across all sequences and datasets

	RC
MoG-PRE	2.42
MoG_C	5.96
MoG_D	3.54
MoG_{Bin}	6.08
$MoG_{RGB\text{-}D}$	6.08
$ViBe_{bin}$	6.92
CL_W	2.79
SOB	4.92
$PBAS$	6.29

Finally, Table 7 reports the overall ranking of all the algorithms using all the sequences of all the datasets. The proposed method, $MoG\text{-}PRE$, has the best ranking score, proving that it has the best average performance for all the different situations.

4 Conclusions

A background modeling strategy that combines depth and color data provided by RGB-D cameras in indoor environments is presente in this paper. The most important contributions of this work are the generation and the efficient combination of different sources of information through a Bayesian framework: the foreground prediction and two background models. The prediction of the foreground between consecutive images is computed. It relies on a novel adaptive block-based foreground modeling, which employs a depth-based dynamic model to robustly predict the temporal evolution of deformable foreground regions. Depth-based and color-based independent per-pixel background models, based on the MoG algorithm, are also included in the combination. The advantage of the proposed approach is that the actual contribution of the depth and color features and their corresponding models for the final foreground segmentation is adapted by the Bayesian framework. Additionally, the combination of the foreground prediction and the per-pixel models allows to include also the spatial information and the local depth characteristics of the foreground regions, thus guaranteeing more compact and accurate detections. The results have demonstrated that the proposed approach outperforms state of the art background modeling strategies based on RGB-D imagery. Indeed, it efficiently tackles strong illumination variations, interference of multiple active RGB-D cameras, non-measured depth data, depth camera noise, and situations of sudden people crowds.

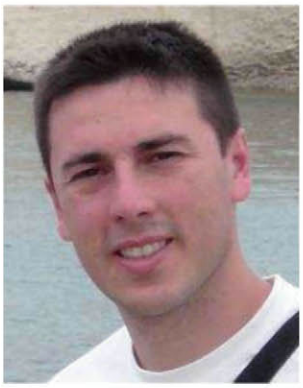
Acknowledgments This work has been partially supported by the Ministerio de Economía y Competitividad of the Spanish Government under the project TEC2010-20412 (Enhanced 3DTV). M. Camplani would like to acknowledge the European Union and the Universidad Politécnica de Madrid (UPM) for supporting his activities through the Marie Curie-Cofund research grant.

References

1. Albiol, A., Albiol, A., Mossi, J., Oliver, J.: Who is who at different cameras: people re-identification using depth cameras. IET Comput. Vision **6**(5), 378–387 (2012)
2. Arulampalam, M., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Trans. Signal Process. **50**(2), 174–188 (2002)
3. Barbosa, I., Cristani, M., Bue, A., Bazzani, L., Murino, V.: Re-identification with RGB-D sensors. In: Computer Vision ECCV 2012. Workshops and Demonstrations. Lecture Notes in Computer Science **7583**, 433–442 (2012)
4. Barnich, O., Van Droogenbroeck, M.: ViBe: a universal background subtraction algorithm for video sequences. IEEE Trans. Image Process. **20**(6), 1709–1724 (2011)
5. del Blanco, C.R., Jaureguizar, F., García, N.: Robust tracking in aerial imagery based on an ego-motion Bayesian model. EURASIP J. Adv. Signal Process. **2010**, 1–19 (2010)
6. del Blanco, C.R., Jaureguizar, F., García, N.: An advanced Bayesian model for the visual tracking of multiple interacting objects. EURASIP J. Adv. Signal Process. **1**, 130 (2011)
7. Bouwmans, T.: Recent advanced statistical background modeling for foreground detection - a systematic survey. Recent Patents Comput. Sci. **4**(3), 147–176 (2011)
8. Bouwmans, T., Baf, F.E.: Background modeling using mixture of gaussians for foreground detection-a survey. Recent Patents Comput. Sci. **3**, 219–237 (2008)
9. Camplani, M., Salgado, L.: Background foreground segmentation with RGB-D Kinect data: an efficient combination of classifiers. J. Vis. Commun. Image Represent. (2013) (in press)
10. Camplani, M., Mantecon, T., Salgado, L.: Accurate depth-color scene modeling for 3D contents generation with low cost depth cameras. In: 2012 19th IEEE International Conference on Image Processing (ICIP), pp 1741–1744 (2012)
11. Camplani, M., Mantecon, T., Salgado, L.: Depth-Color Fusion Strategy for 3D scene modeling with Kinect. IEEE Transactions on Cybernetics (accepted paper) (2013)
12. Cristani, M., Farenzena, M., Bloisi, D., Murino, V.: Background subtraction for automated multisensor surveillance: a comprehensive review. EURASIP J. Adv. Signal Process. **2010**, 1–24 (2010)
13. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification, 2nd edn. Wiley-Interscience, Newyork (2001)

14. Frick, A., Kellner, F., Bartczak, B., Koch, R.: Generation of 3D-TV LDV-content with time-of-flight camera. In: IEEE 3DTV Conference, pp 1–4 (2009)
15. Gordon, G., Darrell, T., Harville, M., Woodfill, J.: Background estimation and removal based on range and color. In: IEEE Computer Society Conference on Computer Vision and. Pattern Recognition **2**, 464 (1999)
16. Goyette, N., Jodoin, P.M., Porikli, F., Konrad, J., Ishwar, P.: Changedetection.net: a new change detection benchmark dataset. In: IEEE computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, pp 1–8 (2012)
17. Han, J., Shao, L., Xu, D., Shotton, J.: Enhanced computer vision with microsoft Kinect Sensor: a review. IEEE transactions on Cybernetics (accepted paper) (2013).
18. Hofmann, M.: Background segmentation with feedback: the pixel-based adaptive segmenter. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 38–43 (2012)
19. KaewTraKulPong, P., Bowden, R.: An improved adaptive background mixture model for real-time tracking with shadow detection. In: European Workshop on Advanced Video Based Surveillance Systems, pp 149–158 (2001)
20. Khoshelham, K., Elberink, S.O.: Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors **12**(2), 1437–1454 (2012)
21. Klare, B., Sarkar, S.: Background subtraction in varying illuminations using an ensemble based on an enlarged feature set. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 66–73 (2009)
22. Kuncheva, L.: Combining pattern classifiers: methods and algorithms. Wiley-Interscience, Newyork (2004)
23. Leens, J., Barnich, O., Piérard, S., Droogenbroeck, M., Wagner, J.M.: Combining color, depth, and motion for video segmentation. In: Computer Vision Systems. Lecture Notes in Computer Science **5815**, 104–113 (2009)
24. Li, L., Huang, W., Gu, I.Y.H., Tian, Q.: Statistical modeling of complex backgrounds for foreground object detection. IEEE Trans. Image Process. **13**(11), 1459–1472 (2004)
25. Maddalena, L., Petrosino, A.: A self-organizing approach to background subtraction for visual surveillance applications. IEEE Trans. Image Process. **17**(7), 1168–1177 (2008)
26. Mastorakis, G., Makris, D.: Fall detection system using kinect's infrared sensor. J. Real-Time Image Process. pp 1–12 (2012). doi:10.1007/s11554-012-0246-9
27. Molina, J., Escudero-Viñolo, M., Signoriello, A., Pardàs, M., Ferrán, C., Bescós, J., Marqués, F., Martínez, J.M.: Real-time user independent hand gesture recognition from time-of-flight camera video using static and dynamic models. Mach. Vis. Appl. **24**(1), 187–204 (2011)
28. Spinello, L., Arras, K.: People detection in rgb-d data. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 3838–3843 (2011)
29. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: IEEE Conference on Computer Vision and, Pattern Recognition, pp 246–252 (1999)
30. Stone, E., Skubic, M.: Evaluation of an inexpensive depth camera for in-home gait assessment. J. Ambient Intell. Smart Environ. **3**(4), 349–361 (2011)
31. Stormer, A., Hofmann, M., Rigoll, G.: Depth gradient based segmentation of overlapping foreground objects in range images. In: IEEE Conference on, Information Fusion, pp 1–4 (2010)
32. Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: principles and practice of background maintenance. In: Proceedings of the Seventh IEEE International Conference on Computer Vision pp 255–261 (1999)

Author Biographies



Massimo Camplani received his MS in electronic engineering in 2006 with honors, both from the Università degli Studi di Cagliari, Italy. In 2010 he received Ph.D. in Electronic and Computer Engineering at the Università degli Studi di Cagliari. Since 2010, he has been a member of the Grupo de Tratamiento de Imágenes (Image Processing Group) of the Universidad Politécnica de Madrid, Spain. In April 2011, he was awarded the Marie Curie-COFUND Grant. His research interests are in the area of computer vision.

Carlos Roberto del Blanco received the Telecommunication Engineering degree and the Ph.D. degree in Telecommunication, both from the Universidad Politécnica de Madrid (UPM), in 2005 and 2011, respectively. Since 2005 he is a member of the Image Processing Group of the UPM. In addition, since 2011 he is a member of the faculty of the E.T.S. Ingenieros de Telecomunicación as Assistant Professor of Signal Theory and Communications at the Department of Signals, Systems, and Communications. His professional interests include signal and image processing, computer vision, pattern recognition, machine learning, and stochastic dynamic models.

Luis Salgado received the Telecommunication Engineer degree in 1990 and the Ph.D. degree in communications with summa cum laude in 1998, both from the E.T.S.I. Telecomunicación, Universidad Politécnica de Madrid (UPM), Spain. Since 1990, he has been a member of the Image Processing Group (GTI) of the UPM. Since 1996 he has been a member of the faculty of the UPM, formerly as a teaching assistant, and currently as associate professor (tenure in 2001) of Signal Theory and Communications in the Dept. of Signals, Systems, and Communications. From September 2012 he has also joined the Universidad Autónoma de Madrid as associate professor at the Escuela Politécnica Superior (VPULab Group, Dept. of Electronics and Communications Technology). He is associate editor of the Journal of Real-Time Image Processing, has been member of the Scientific and Program Committees

of several international conferences and has been auditor and evaluator of European research programs since 2002. He has participated in many national and international research projects, and his professional interests include video analysis, processing and coding.

Fernando Jaureguizar received the Telecommunication Engineering degree and the Ph.D. degree in Telecommunication, both from the Universidad Politécnica de Madrid (UPM), in 1987 and 1994, respectively. Since 1987 he is a member of the Image Processing Group of the UPM. In addition, since 1991 he is a member of the faculty of the E.T.S. Ingenieros de Telecomunicación at UPM, and since 1995 he is an Associate Professor of Signal Theory and Communications at the Department of Signals, Systems, and Communications. His professional interests include digital image processing, video coding, 3DTV, computer vision, and design and development of multimedia communications systems. He has been actively involved in European projects (Eureka, ACTS and IST) and national projects in Spain.

Narciso García received Telecommunication Engineering degree and the Ph.D. degree in Telecommunication, both from the Universidad Politécnica de Madrid (UPM), in 1976 (Spanish National Graduation Award) and 1983 (Doctoral Graduation Award), respectively. Since 1977 he is a member of the faculty of the UPM, where is currently Professor of Signal Theory and Communications. He leads the Image Processing Group of the UPM. He was Coordinator of the Spanish Evaluation Agency from 1990 to 1992 and evaluator, reviewer, and auditor of European programs since 1990. His professional and research interests are in the areas of digital image and video compression and of computer vision.