Multi-sensor background subtraction by fusing multiple region-based probabilistic classifiers

Massimo Camplani , Carlos R. del Blanco , Luis Salgado , Fernando Jaureguizar , Narciso García

A B S T R A C T

In the recent years, the computer vision community has shown great interest on depth-based applications thanks to the performance and flexibility of the new generation of RGB-D imagery. In this paper, we present an efficient background subtraction algorithm based on the fusion of multiple region-based classifiers that processes depth and color data provided by RGB-D cameras. Foreground objects are detected by combining a region-based foreground prediction (based on depth data) with different background models (based on a Mixture of Gaussian algorithm) providing color and depth descriptions of the scene at pixel and region level. The information given by these modules is fused in a mixture of experts fashion to improve the foreground detection accuracy. The main contributions of the paper are the region-based models of both background and foreground, built from the depth and color data. The obtained results using different database sequences demonstrate that the proposed approach leads to a higher detection accuracy with respect to existing state-of-the-art techniques.

1. Introduction

Recently, depth data processing and analysis have achieved a great importance in many computer vision applications. In particular, thanks to the presence of low-cost depth cameras in the market, such as the Microsoft Kinect that provides both color and depth information at high frame rates, several applications have emerged from the computer vision research community that make use of this rich information. One of the most important research areas in which the depth data has been successfully employed is the human motion analysis, as presented by Chen et al. (2013): the depth data is used to identify and segment human users in the scene in order to accurately track their body parts. These data are then processed and used in controller-free human–computer interaction systems; particular attention has been paid to gesture recognition systems such as the one presented by Mahbub et al. (2013). The use of RGB-D imagery has been also positively applied in different computer vision tasks and applications for indoor environments, such as the robot-based application presented by Doisy et al. (2012), the video surveillance system proposed by Clapés et al. (2013), the smart environment for ambient assisted living presented by Stone and Skubic (2011), and the human detection algorithm proposed by Spinello and Arras (2011).

In applications such as indoor video surveillance or human computer interaction, the information provided by the depth data helps to separate the moving objects from the static scene for further analysis and processing. Hence, robust background subtraction algorithms, based on the fusion of color and depth data, are required to improve the performance of depth-based applications.

Background subtraction is a key processing step of many computer vision applications. It aims at separating the moving objects in the scene (that constitute the foreground) from a robust model of the static environment (the background). As described by Cristani et al. (2010), the performance of color-based algorithms highly depend on the background model initialization, background multimodality, and it deteriorates with the presence of color camouflage, illumination variations, and cast shadows. Robustness against the latter issues can be achieved incorporating depth data provided by low-cost depth cameras to the model. However, depth data presents several problems that negatively affect depth-based background modeling algorithms. In particular, object silhouettes are heavily affected by the high level of noise at object boundaries, as shown by Camplani et al. (2012). Furthermore, depth data cannot be estimated for all the image pixels due to occlusions, reflections, or out-of-range points, as presented by Camplani et al. (2013) (we will call these data as non measured pixels (*nmd*)). Moreover, depth measurements provided by structured light sensors, such as the Microsoft Kinect, are affected by noise process that follows a quadratic relationship with the measured depth value, as presented by Khoshelham and Elberink (2012).

Although different background subtraction techniques have been presented in the literature, there are very few approaches that propose a fusion of both color and depth data to improve the algorithm performance. For more details, see the reviews presented by Cristani et al. (2010) and by Bouwmans (2011) about background subtraction, and the very recent overview about advances in RGB-D based applications proposed by Han et al. (2013).

Gordon et al. (1999) presented one of the first works based on the fusion of color and depth data obtained from a stereo device. This work is based on the Mixture of Gaussians (MoG) algorithm proposed by Stauffer and Grimson (1999). A per-pixel background model is built using a four dimensional mixture of Gaussian distribution: one component is the depth, and the other three are color features (YUV color space is employed). Depth and color features are assumed independent.

The MoG algorithm has been also used by Stormer et al. (2010) to combine depth and infrared data. Two independent per-pixel background models are built, and pixels are classified as foreground when both models agree, otherwise the pixels are classified as background. However, the performance of this approach is severely affected by the misclassification errors from each model.

Leens et al. (2009) propose combining a color camera with a Time-of-Flight (ToF) camera for video segmentation. As in previously mentioned approaches, color and depth data are assumed to be independent, and the Vibe algorithm (presented by Barnich and Van Droogenbroeck (2011)) is applied to obtain the foreground masks, which are combined with logical operations, and filtered with morphological operators.

Recently, in the Microsoft Kinect based surveillance system proposed by Clapés et al. (2013), a per pixel background subtraction technique is presented. The authors propose a background model based on a four dimensional Gaussian distribution (using color and depth features). This approach is quite limited since it cannot manage multimodal backgrounds, and does not address the depth-data noise issues associated to the Kinect.

In the gesture recognition system presented by Mahbub et al. (2013), the foreground silhouette objects are extracted by applying a threshold approach proposed by Otsu (1979) to the depth data. The results reported show good performance in very controlled environments characterized by a constant background, and with the additional restriction that there can be only a single user in the scene who must be well separated (in depth) from the background.

Camplani and Salgado (2013) propose a per-pixel background modeling approach that fuses different statistical classifiers based on depth and color data by means of a weighted average combiner that takes into account the characteristics of depth and color data. A mixture of Gaussian distribution is used to model the background pixels, and a uniform distribution is used for the modeling of the foreground.

In this paper, we propose an innovative background subtraction algorithm for processing multi-sensor data provided by RGB-D cameras in indoor environments. The proposed approach fuses multiple region-based classifiers in a mixture of experts fashion to improve the final foreground detection performance. It is based on multiple background models that provide a description at region and pixel level by considering the color and depth features. These models are based on the Mixture of Gaussian algorithm. Background regions are identified by independently applying Mean Shift, proposed by Comaniciu and Meer (2002), on depth and color data. Moreover, we provide a region-based foreground prediction that relies on depth data. In particular, a depth-histogram appearance model of the foreground is combined with two spatial and depth-based dynamic models to predict the expected depth and position of the foreground regions. Data from the background models and the foreground prediction are then fused in a mixture of experts system that efficiently combines the contribution of the color and depth features to render the foreground segmentation. The main contributions of the proposed approach are: the combination of the pixel-based and region-based background models that fuse color and depth data; the foreground prediction scheme; and the region-based foreground model. Results using different publicly available datasets demonstrate that the proposed technique efficiently tackles strong illumination variations, interferences due to the existence of multiple active RGB-D cameras, depth data noise, non-measured depth data, and the presence of sudden crowds.

The rest of the paper is structured as follows: in Section 2, the proposed strategy is presented; results are shown in Section 3. Lastly, conclusions are drawn in Section 4.

2. Multi-sensor background subtraction algorithm

The scheme of the proposed multi-sensor background algorithm is presented in Fig. 1. Mean shift (*MShift* block) is applied to the depth and color data, D_t and C_t, to obtain the corresponding segmented maps, $MS-D_t$ and $MS-C_t$. Segmentation maps and actual depth and color information are used by the background modeling block (*BgMOD* in Fig. 1) to build four independent background models by considering the temporal evolution of the depth and color data at both pixel and region level. In parallel, the *RegPRED* block computes a prediction of the foreground and background probability maps for the current time instant (p_{fg} and p_{bg} in Fig. 1) using the previous depth data and the segmented foreground regions (Fg_{t-1} in Fig. 1) and the available depth-based background model (Bg_t in Fig. 1). These probability maps play the role of prior foreground/background probabilities for each image pixel, whereas the four background models are used to obtain the foreground/background likelihood maps (L in Fig. 1). Prior probabilities and likelihoods are combined to estimate the posterior probability of each class using a Bayesian perspective. Finally, the different posterior probabilities are fused together in a mixture of experts fashion by the *MoE* block to obtain a more reliable estimation of the foreground regions. In particular, a weighted average scheme that takes into account depth discontinuities and the non-measured depth (*nmd*) pixels distribution is used in the combination of the posterior probabilities. In the following sections, further details on the blocks that constitute the proposed algorithm are given.

2.1. Pixel-based and region-based background modeling

Four models are computed for the scene background that describe the static scene at pixel and region level by considering independently depth and color features.

Two independent per pixel models, depth-based and color-based, are iteratively built and updated using the Mixture of Gaussian algorithm (MoG) presented by Stauffer and Grimson (1999). This popular algorithm uses a parametric model based on a mixture of Gaussians to represent the statistical distribution of each image pixel. The main advantages of this approach are its capability to handle multimodal backgrounds and gradual changes of the scene. Distribution parameters are iteratively updated with an online version of the Expectation Maximization algorithm.

The MoG is a two-step algorithm: in the first step, it is tested whether or not every incoming pixel value belongs to the background model, and in the second step, the model parameters are recursively updated. As reported by Zivkovic and van der Heijden (2006), the mixture of Gaussian distribution models at the same time the probability that one pixel belongs to the background and to the foreground. In particular, the most probable Gaussians,

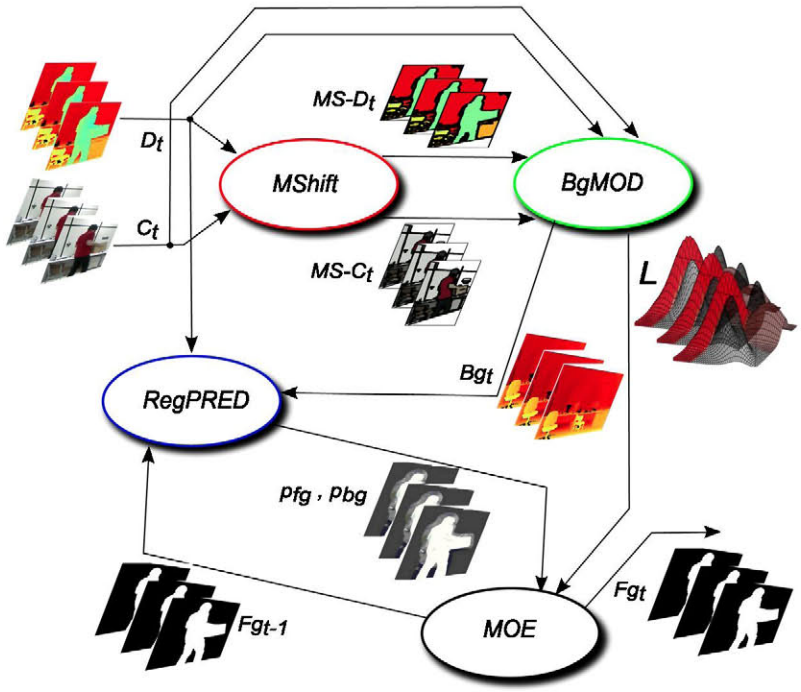

Fig. 1. Scheme of the proposed multi-sensor background subtraction approach. D_t and C_t are respectively the depth and the color data. Previous foreground detection and the depth-based background model are indicated with Fg_{t-1} and Bg_t. The likelihood probabilities estimated by *BgMOD* are indicated with *L*, and probabilities of the foreground and background regions predicted by *RegPRED* are indicated with p_{fg} and p_{bg}.

characterized by a high weight and a low variance, are considered as the background ones. The Gaussians are ranked after each iteration by considering the factor $r = \omega/\sigma$, where ω is the weight of a Gaussian and σ its variance. To estimate if a pixel value belongs to one distribution, the Mahalanobis distance is used. If no proper match is found, the least probable Gaussian is substituted by a new one characterized by a high variance and a mean value equal to the pixel value. The parameters for the matched distribution are iteratively updated as follows:

$$\begin{aligned}
\omega_{i,t+1} &= \omega_{i,t}(1-\alpha) + \alpha * \ \Gamma \\
\rho &= \alpha\eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \\
\mu_{i,t+1} &= \mu_{i,t}(1-\rho) + \rho X_t \\
\sigma^2_{i,t+1} &= \sigma^2_{i,t}(1-\rho) + \rho\left(X_{t+1} - \mu_{i,t+1}\right)^2
\end{aligned} \tag{1}$$

where μ_i is the mean of the *ith* Gaussian distribution, and α is the learning rate, a parameter that determines the adaptation to changes in the scene and the speed of incorporation of foreground objects to the background model. For the unmatched Gaussians, all the parameters remain unchanged, except their weight that is updated with $\Gamma = 0$ in (1). A complete review about the MoG algorithm performance and its several modifications can be found in the survey presented by Bouwmans and Baf (2008).

In this work, we propose several modifications to the original MoG algorithm to improve the detection performance. Instead of using a fixed learning rate value to update the distribution parameters, we propose using a variable learning rate, as proposed by KaewTraKulPong and Bowden (2001). Its value is decreased at each iteration until a minimum fixed value of α is reached, thus limiting the absorption of moving objects to the background model at the beginning of the sequence.

As far as the per-pixel color-based MoG model and the corresponding classifier (hereafter MoG_C) is concerned, it is necessary to reduce the effect of sudden changes of illumination that can lead to wrong pixel classification. In this case, we propose to use the frame-level control strategy proposed by Toyama et al. (1999), where the fraction of the pixels detected as foreground is computed and compared with a predefined threshold. If the computed fraction exceeds this threshold, the MoG_C Gaussian parameters and the corresponding learning rate are re-initialized. This frame-level control is important if it is not possible to control the acquisition settings of the RGB-D camera (as it is the case of the Microsoft Kinect used in our experiments), resulting in sudden changes of the luminance component of the images.

Regarding the per-pixel depth-based model and the corresponding classifier (hereafter MoG_D), we modify the original MoG algorithm to reduce the effect of the distance-dependent noise, as proposed by Camplani and Salgado (2013). As already mentioned in Section 1, there is a quadratic dependency between the measured depth value, d, and the standard deviation of the noise, which affects the measurement, $\sigma_{noise}(d)$. Thus, the larger the sensor-object distance, the higher the noise level affecting the object depth measurements. This can bias the ranking of the Gaussian distributions, and as a consequence to increase the depth-based pixel misclassification rate. To mitigate the impact of the distance-dependent noise on the classification process, the ranking parameters is normalized with the factor $\sigma_{noise}(\mu_i)$ that is selected according to the quadratic distance-noise relationship presented by Khoshelham and Elberink (2012).

At region level we propose other two independent models (also based on the MoG), which take into account the spatial characteristics of the background. One model classifies pixels by considering the depth data, and the other one by considering the color data. We define as MoG_{RD} and MoG_{RC} the classifiers based respectively on the region depth-based and region color-based models.

Jodoin et al. (2007) demonstrated that the distribution of neighbor pixel values (spatial information) can be successfully used to build a robust background model of the analyzed scene, which has similar characteristics to the one obtained by considering the per-pixel temporal evolution. The authors demonstrate the validity of this assumption by incorporating the spatial information in

popular background subtraction algorithms based on both non-parametric and parametric approaches. In their paper Jodoin et al. (2007), propose to use a square neighborhood for each pixel and fit on this area either a Gaussian distribution or a mixture of Gaussians distribution. Moreover, the model parameters are updated in the spatial version of the parametric algorithm, as originally proposed by Stauffer and Grimson (1999), but including the contribution of all the neighbor pixels.

Following these ideas, in our proposed region based approach, the two classifiers MoG_{RD} and MoG_{RC} are built taking into account the spatial distribution of the depth and color, respectively. However, we modify the neighborhood definition with respect to the original paper of Jodoin et al. (2007) to take into account the spatial coherency. Instead of a fixed size neighborhood, we propose using the regions resulting from the mean-shift segmentation, regions whose pixels share similar properties in depth or color. In particular, for each pixel we built two MoG-like classifiers (MoG_{RD} and MoG_{RC}). The principal mode of the corresponding image region/cluster (identified by the mean-shift technique in depth and color, respectively) is used to compute the mean and variance values of the Gaussian distribution. MoG_{RD} and MoG_{RC} parameters are updated as in (1), and similarly to the pixel-level models, the frame level control and the ranking parameters normalization are applied.

Therefore, we built four different classifiers based on the MoG algorithm for each pixel: two per-pixel classifiers obtained by processing the temporal evolution of the pixel color and depth values (based on C and D), and two region-based classifiers (based on $MS-C$ and $MS-D$) obtained by processing the temporal evolution of the regions that have been identified by the mean shift algorithm.

2.2. Foreground prediction based on depth histograms

A depth-based appearance model is combined with two spatial and depth-based dynamic models to predict the expected depth and position of the foreground objects along the time. The appearance model uses a set of depth histograms to encode the foreground depth information. The depth-based dynamic model predicts the expected depth of the foreground regions between consecutive images. The spatial dynamic model estimates the spatio-temporal evolution of the foreground regions. All these models are used to compute a foreground probability density function (pdf), p_{fg}, which complements the information given by the background model block (*BgMOD* in Fig. 1) to improve the foreground/background segmentation.

The estimation of p_{fg} can be split into 5 stages: (1) semi-dense computation of foreground depth histograms, (2) clustering of foreground depth histograms, (3) depth evolution of foreground regions, (4) spatial evolution of foreground regions, and (5) final estimation of the foreground pdf.

During the first stage, a semi-dense computation of foreground depth histograms takes place, starts by uniformly sampling the foreground regions indicated by Fg_{t-1} (the foreground regions obtained at the previous time step). As a result, $S_{fg}=\{x_i\in Z^2 | i=1,\ldots,N_s\}$ is obtained, which is a subset of spatial coordinates of the foreground regions, where N_s is the number of spatial coordinates in the set S_{fg}.Then, a depth histogram is computed for each element of S_{fg} by considering the data inside a square region defined by $\{x_i, l\}$, where x_i is the center and l is the half-side of the square.

Along the second stage, the clustering of foreground depth histograms is performed. It reduces the number of foreground depth histograms to speed up the computation of the algorithm: depth histograms computed from overlapped or just close image regions are usually very similar (high redundancy). The k-means clustering strategy (see the clustering section of Kuncheva (2004)) can be applied to obtain a set of k representative depth histograms for the foreground. The obtained k cluster centroids are normalized to one to ensure that they are rigorously histograms. This set of k histograms, indicated with S_h, represents a depth-based appearance model of the foreground for the time step $t-1$.

In the third stage, a depth-based dynamic model predicts the depth evolution of the foreground regions between time steps. The depth of the foreground regions changes because of the own motion of the objects that form the foreground. This means that the depth histograms of the foreground regions will be different between consecutive frames. A constant velocity model is used to predict the new depth values of the foreground. This dynamic model is directly applied over the set of depth histograms S_h for efficiency purposes, instead of being applied over the raw depth values, since a shift in the depth values of a determined region is equivalent to a shift in the corresponding depth histogram. In addition, a linear interpolation technique is used to compute the final set of predicted depth histograms S'_h, since the predicted displacement in depth is not in general a multiple of the bin width of the histograms.

The fourth stage, spatial evolution of foreground regions, uses a spatial dynamic model to predict the location of the foreground regions between consecutive time steps. Likewise the previous stage, a constant velocity model is used for the spatial prediction of the foreground. This set of coordinates that determine the predicted spatial locations are called S'_l.

As a result of the third and fourth stages, a set of foreground spatial locations S'_l and other of foreground depth histograms S'_h are obtained for the current time step.

The last stage, final estimation of the foreground pdf (p_{fg}), evaluates the probability that one image pixel is foreground. For this purpose, the predicted set of candidate foreground locations S'_l, the predicted foreground depth histograms S'_h, and the depth histograms obtained from the current depth image D_t are used. The Bhattacharyya distance is used in the computation of p_{fg} to evaluate the similarity between one predicted depth histogram $h'\in S'_h$ and one candidate depth histogram h, computed from a region of D_t defined by $\{x'_i, l\}$, where $x'_i\in S'_l$ is one of the potential foreground spatial locations in the current time step. The mathematical expression of the Bhattacharyya distance is

$$b_d(h',h)=\sqrt{1-b_c(h',h)}, \tag{2}$$

where b_c is the Bhattacharyya coefficient given by

$$b_c(h',h)=\sum_i \sqrt{h'(i)h(i)}, \tag{3}$$

where i is the index to iterate on the histogram bins.

The foreground pdf p_{fg} for one pixel whose spatial coordinates x'_i belong to the set S'_l is estimated as

$$p_{fg}=\max_{h'\in S'_h}\left(N\left(b_d(h',h);0,\sigma_f^2\right)\right), \tag{4}$$

where σ_f^2 is the variance that models the expected uncertainty in the prediction of the foreground depth histograms. Therefore, the foreground pdf depends on the maximum similarity value between one candidate histogram and the set of the predicted depth histograms, S'_h. The advantage of encoding the foreground appearance with this multiple depth histogram model is that it provides good results for deformable moving objects such as humans.

2.2.1. Background pdf based on depth histograms

The background pdf p_{bg} could be just computed as $p_{bg}=1-p_{fg}$. However, this simple approach is prone to errors in real situations. The reason is that the segmented foreground Fg_{t-1} used to predict

p_{fg} has (almost) always false positives (pixels that actually belong to the background). This implies that some background regions could have a high p_{fg}, since one (or more) of the predicted depth histograms in S'_h has been computed with the false positive pixels corresponding to the background. Even more, this situation degenerates along time because of the feedback, continuously increasing the number of errors, and thus making useless the computed p_{fg} and p_{bg}. To solve this problem, the background pdf p_{bg} is explicit and independently computed, counteracting the effect of the false positives in the segmented foreground regions. The idea is as follows. Let us consider first the case that both probabilities are independently computed. If some background regions have been erroneously classified as foreground in the previous time step, both probabilities p_{fg} and p_{bg} could be high for some background region in the current image, since both appearance models, background and foreground, include image regions from the actual background. However, both probabilities would have approximately the same value after their normalization (their sum must be one). As a consequence, the information given by the foreground prediction module is not decisive, and the final classification as foreground or background will fall on the information given by the other modules (the pixel-based background models). On the other hand, consider the case that only p_{fg} is explicitly computed, and then $p_{bg} = 1 - p_{fg}$. In this case, if p_{fg} is erroneously high for some background region, the corresponding p_{bg} will be low. As a consequence, that region will tend to be misclassified as foreground, in spite of the information given by the other modules.

The computation of p_{bg} is similar to the computation of p_{fg} explained in the previous section. The key difference is that the background regions are assumed to be static, and therefore the depth histogram used to model the background appearance of a region is computed from the depth background model, Bg_t, provided by the *BgMOD* block (see Section 2.1). Thus, the background pdf p_{bg} for the spatial coordinate $x'_i \in S'_l$ (where S'_l is the same set as in the previous section) is estimated as

$$p_{bg} = N(b_d(h'', h); 0, \sigma_b^2), \tag{5}$$

where h'' is the depth histogram computed from the depth background model, Bg_t, in the square region defined by $\{x_i, l\}$. And σ_b^2 is the variance that models the expected uncertainty in the prediction of the depth histogram for the background.

2.3. Mixture of experts

As proposed by Aach and Kaup (1995), the challenging problem of background and foreground segmentation can be viewed as a classification problem where for each pixel a class label has to be assigned. In the previous sections, we have introduced four different classifiers based on the analysis of different features: spatial and temporal evolution of depth and color data. Our objective is to combine these classifiers and the predictions provided by the *RegPRED* block to improve the final classification performance of our system.

The combination of classifiers, also referenced as mixture of experts, is a popular and efficient strategy employed to solve pattern recognition tasks. It is based on the idea that a set of different simple classifiers can guarantee more robust performance than a single complex classifier. In particular, in our case the mixture of experts is useful to efficiently solve a data fusion problem, as underlined by Polikar (2006), where the data provided by different sources has to be combined to obtain the final classification. In these cases, poor performance are generally obtained if a single classifier is used to learn the information contained in all the data. A complete review and analysis about mixture of experts can be found in the book presented by Kuncheva (2004) and the work proposed by Polikar (2006).

As mentioned in Section 2.1, the MoG based classifiers allow to calculate the likelihood (indicated generally as L in the Fig. 1) that a pixel belongs to the background class ω_{bg} or to the foreground class ω_{fg}. In the proposed approach, we combine the estimated likelihood (one for each model), with the prediction provided by the *RegPRED* block. In particular, we consider the predicted probabilities (p_{fg} and p_{bg}) as the class prior probabilities, and then we compute the posterior probability that the measured data belong to one of the two classes by applying the Bayes' rule. The obtained posterior probabilities are then fused by the *MoE* to obtain the final classification. The estimation of posterior probabilities does not depend on the sequence, and they are estimated in an online fashion for each processed frame following the evolution of the adaptive likelihood models and the predicted prior values that consider the depth evolution of the foreground regions. In the following paragraphs, we describe the main features of the *MoE* module.

As proposed by Kuncheva (2004), we calculate for each pixel the decision profile DP containing the support of each classifier to the hypothesis that the measured data belongs to one of the two classes. Considering the pixel s at position (x, y), and the corresponding measured data $\mathbf{x}_s$, the decision profile is:

$$DP(\mathbf{x}_s) = \begin{bmatrix} d_{C,bg} & d_{C,fg} \\ d_{D,bg} & d_{D,fg} \\ d_{RC,bg} & d_{RC,fg} \\ d_{RD,bg} & d_{RD,fg} \end{bmatrix} \tag{6}$$

where each row represents the estimated posterior probabilities obtained by combining the likelihoods estimated by *BgMOD* and the prior probabilities estimated by *RegPRED*. Each column of the decision profile represents the overall support $M(\mathbf{x}_s)$ of the set of classifiers to one of the classes. The information contained in DP can be used to estimate the overall support for all the classes, and also to assign class labels to image pixels depending on the class that has the greatest value in $M(\mathbf{x}_s)$.

Two main techniques for the combination of the data contained in $DP(\mathbf{x}_s)$ have been presented in the literature: *class conscious* approaches, in which the overall support of the classes are not fused together; and *class indifferent* approaches, where the combination of the values of M represents new features that are processed in a successive step by other classifiers in order to obtain the final classification. In our approach, we select a *class conscious* approach because it guarantees low computational requirements and does not provide additional parameters to the system, since the overall supports are obtained with arithmetic operations.

Typical combination rules used in *class conscious* approaches are: average, median, maximum, etc. (see the review presented by Kuncheva (2004, Ch. 5) for more details). An example of these approaches in the field of background subtraction can be found in the work proposed by Klare and Sarkar (2009) where a simple average is used: the authors propose a mixture of experts system based on 13 classifiers, that process 13 different visual cues. The simple average approach gives to all the classifiers the same influence to the final classification. In our approach we propose a weighted average as it is able to extract different information from the different features (regions, color, depth) and to efficiently adapt the *contribution* of each classifier to the final classification. For the j_{th} class the overall support $M_j(\mathbf{x}_s)$ is estimated as:

$$M_j(\mathbf{x}_s) = \sum_{i=1}^{4} W_i(\mathbf{x}_s) d_{i,j} \tag{7}$$

In our case we have selected the weight for each classifier (W_i) as a function of the input $\mathbf{x}_s$ in order to increase the support of the most

reliable classifier according to the characteristics of specific image regions. It has to be noticed that, once the weights are estimated, they are normalized to sum one as suggested by Kuncheva (2004). In the following we refer, without loss of generality, to W_C and W_D as the weights assigned to the color based and depth based classifiers for both based on region or pixel features. It is worth noting that for each processed frame, the parameters of the *DP* and the weights presented in Eq. (7) are re-estimated. In particular, the estimated posterior probabilities used in the decision profile are calculated using the adaptive background and foreground models presented in the previous sections. Their values do not depend to any particular sequence and the initial values of the likelihoods can be estimated with just one frame as proposed by Jodoin et al. (2007). The adaptive weights selection strategy is detailed in the following sections.

When the depth measurement (non measured depth *nmd* pixels) or the depth-based background model is not available, only the color based classifiers can be used for the final pixel classification. For this reason, in these cases the weight (W_D) are set to zero, on the contrary the color-based classifier weight (W_C) is set to one.

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 Camplani and Salgado (2013). This is due to the fact that depth data guarantee generally compact detection of moving object regions except 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.

We estimate an edge-closeness probability P^e_G for each pixel, calculated as a function of the distance between the pixel and the closest edge weighted with a Gaussian function. By analyzing the depth and the color data the two edge-closeness probability (P^e_D) and (P^e_C) are obtained; with these values the global edge-closeness probability is calculated such that $P^e_G = P^e_C * P^e_D$. The value of P^e_G is high for those regions for which the color edges correspond to depth edges; in these regions it is necessary to assign a higher value to W_C. 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}, respectively 0.1 and 0.9 in our implementation), in this way it is guaranteed that both classifiers contribute to the final classification.

Moreover, for those pixels for which the corresponding background model is characterized frequently by the presence of *nmd* pixels we propose to modify the weight W_D for the depth based classifier. Let us consider a region that contains several *nmd* pixels for which the depth based model is not available. If a moving object passes throughout this region, the depth-based model is initialized with these depth values that do not correspond to the real background object, thus leading to errors in further pixels classification. For this reason, the value of W_D needs to be scaled by the temporal consistency score *tc* in order to reduce the influence of the depth based classifiers in the final classification of the pixels in these regions. This parameter is estimated as:

$$tc = \frac{1}{1 + e^{\beta \#Hit}} \tag{8}$$

where $\#Hit$ is a counter that is incremented (or decremented) when a valid (invalid) depth measurement is obtained. Hit_{max} is the value for the counter $\#Hit$ for which the *tc* value is equal to one, in this case, the depth based background model is completely reliable. The value of the parameter β is selected according to the chosen value of $\#Hit$. Thanks to the use of *tc*, it is possible to give a different weight to the depth based classifiers relative to those regions negatively characterized by the presence of *nmd* pixels.

3. Results

Results using the proposed strategy and other state-of-the-art algorithms are presented in this section. Two different databases with indoor sequences are employed to test the performance of the algorithms. The first one proposed by Spinello and Arras (2011) is composed by different indoor sequences acquired by an array of three Kinect devices with partially overlapped views. These sequences are characterized by changes in illumination, mutual interference of the Kinect devices (they are active sensors that emit structured light), large regions containing *nmd* pixels (due to reflections and out-of-range data), and the existence of crowded scenes. The ground truth of the database has been manually labeled for this paper, since the original database lacks of region-based ground truth. The ground truth has been generated for every five frames.

From the dataset proposed by Camplani and Salgado (2013), we select the *GenSeq* sequence that represents an indoor environment where only one person is moving in the scene.

Well-known metrics have been used to assess the performance of the algorithms: False Positive rate (FP) (fraction of the background pixels that are marked as foreground), False Negative rate (FN) (fraction of foreground pixels that are marked as background), Total Error (TE) (total number of misclassified pixels normalized with respect to the image size), and a similarity measure (*S*) defined by Li et al. (2004): this metric fuses together the concepts of FP and FN, in such a way that its value is close to 1 if the foreground regions detected are similar to the ground truth ones, and close to 0 if they are very different.

The previous metrics have been used to evaluate state of the art algorithms. Specifically, the selected algorithms are: (1) the binary combination of foreground masks obtained by two MoG based modules (MOG_{Bin}) proposed by Stormer et al. (2010), (2) the binary combination proposed by Leens et al. (2009) ($Vibe_{Bin}$), (3) a RGB-D Mixture of Gaussian algorithm presented by Gordon et al. (1999) (MOG_{RGB-D}) and 4) the pixel-wise combination of classifiers (CL_W) presented by Camplani and Salgado (2013). In addition, we also report the results obtained using only the four mixture of Gaussian modules employed in our system, i.e., discarding the rest of modules that are the main novelties of the paper.

In Fig. 2, details of the data provided by the RGB-D camera and the results of the background subtraction algorithms for the first sequence (Lobby1) of the database proposed by Spinello and Arras (2011) are reported. As it can be noticed, large area containing *nmd* pixels (marked in black) are present in the depth map (see Fig. 2(b)). The proposed approach guarantees an accurate foreground object detection for people moving in the first plane, but also for people moving in the stairs on the back (see Fig. 2(d)). As far as the individual classifiers are concerned, it can be noticed that the color based classifiers (see Fig. 2(e) and (f)) allow to accurately detect the moving objects in the stairs, but due to color camouflage, lead to a fragmented foreground region corresponding to the human moving on the front. On the contrary, the depth based classifiers (reported in Fig. 2(g) and (h)) allow to obtain a more compact silhouette but fail to properly detect the moving objects in regions located out of the device range. It is worth noting that the proposed approach allows to combine efficiently these models and the foreground prediction leading to an improvement of the final foreground segmentation. The algorithm CL_W allows to efficiently detect the moving object in the back stairs, but the foreground object segmented results more fragmented (Fig. 2(i)). The algorithm MoG_{RGB-D} is characterized by a higher number of false positive detections (Fig. 2(l)).

The algorithms detection accuracy is reported in Table 1; it is worth noting that the results obtained with the proposed approach

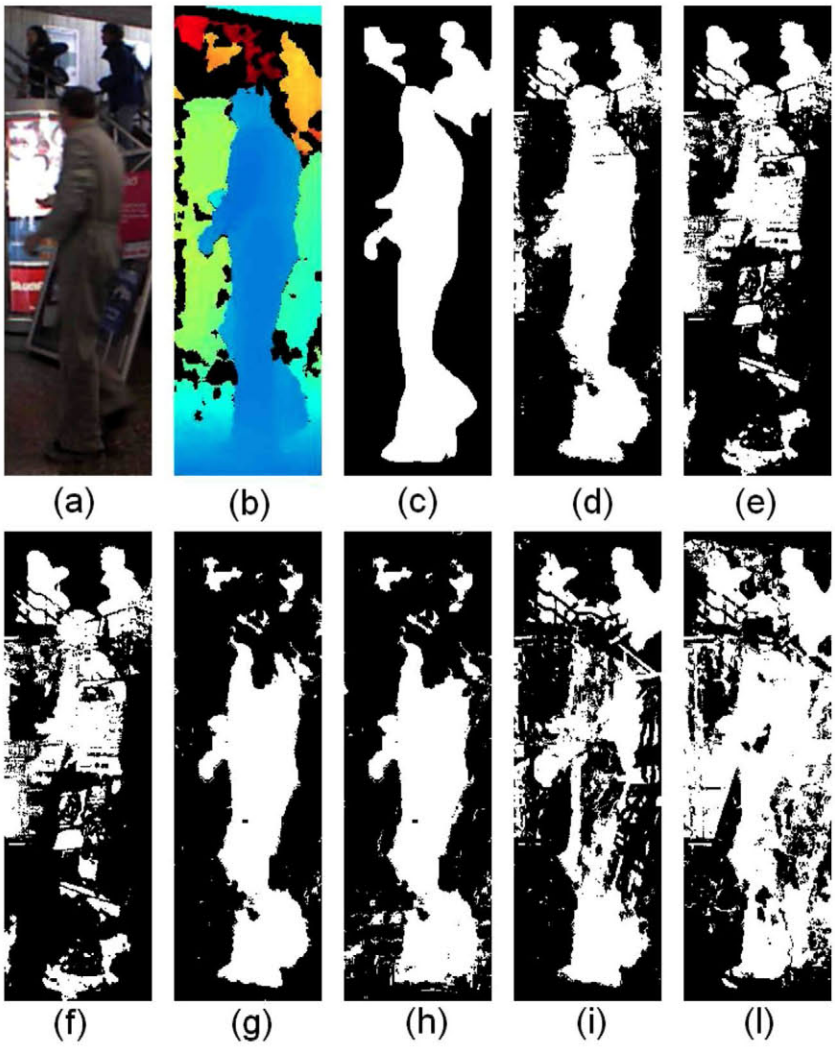

Fig. 2. Lobby1 sequence frame 340: (a) color data, (b) depth data (*nmd* pixels in black), (c) ground truth, (d) proposed $MoG - RegPRE$, (e) MoG_C, (f) MoG_{RC}, (g) MoG_D, (h) MoG_{RD}, (i) CL_W proposed by Camplani and Salgado (2013), (l) MoG_{RGB-D} proposed by Gordon et al. (1999).

(in this and other tables) have been highlighted with bold font. As it can be noticed, the proposed strategy allows to obtain the best results in terms of the similarity measure S and leads to a very low value for TE. It allows to obtain a good tradeoff between the results of the independent classifiers: the value of FN is dramatically reduced with respect to the color based classifiers and at the same time it guarantees a low value for FN. It is worth noting that the other state of the arts techniques that are used during these tests, do not allow to obtain comparable results except for CL_W. In fact, this approach guarantees a similar value for S, however it is affected by a higher fraction of FP, this is due to the fact that this algorithm does not consider the region-based information of the foreground and background objects, thus leading to less compact detections.

Table 1
Detection accuracy obtained by analyzing the Lobby1 sequence.

	TE	FN	FP	S
$MoG - RegPRE$	**5.51**	**20.49**	**3.36**	**0.50**
MoG_C	11.43	55.91	5.04	0.28
MoG_D	5.42	31.41	1.69	0.41
MoG_{RC}	12.26	56.97	5.83	0.25
MoG_{RD}	7.93	31.40	4.55	0.33
MoG_{Bin}	12.00	66.73	4.13	0.18
MoG_{RGB-D}	25.26	10.61	27.36	0.26
$ViBe_{bin}$	14.99	45.87	10.55	0.23
CL_W	7.72	15.45	6.61	0.45

Fig. 3 shows a detail of depth and color data ((a) and (b)) of the second sequence (Lobby2) of the database proposed by Spinello and Arras (2011) and the foreground detection obtained with the different algorithms. Let us consider the color based classifiers (Fig. 3(e) and (f)), as it can be noticed the legs of the human are not well segmented; in particular, there is one squared hole in the foreground regions, due to the color camouflage, and the detection is affected by strong reflections in the floor due to a door opening in the back wall. On the contrary the illumination conditions do not affect the depth based classifiers (Fig. 3(g) and (h)); however, they are affected by the presence of large areas with *nmd* pixels that cause a poor segmentation of the legs. Finally, it can be noticed how the noise that affects depth data at object boundaries reduces the accuracy of the segmented object boundaries (e.g., in the human head). The proposed approach $MoG - RegPRE$ allows to improve the final foreground detection accuracy, by recovering the detection of the human legs thanks to the foreground region analysis and the temporal consistency score *tc*. Moreover, more refined object boundaries are obtained. Refined object boundaries are obtained also with CL_W (Fig. 3(i)); however, the absence of the analysis of foreground regions and of the temporal score do not allow to completely recover the human legs. It has to be noticed that the foreground regions obtained by MoG_{RGB-D} (Fig. 3(l)) are in this

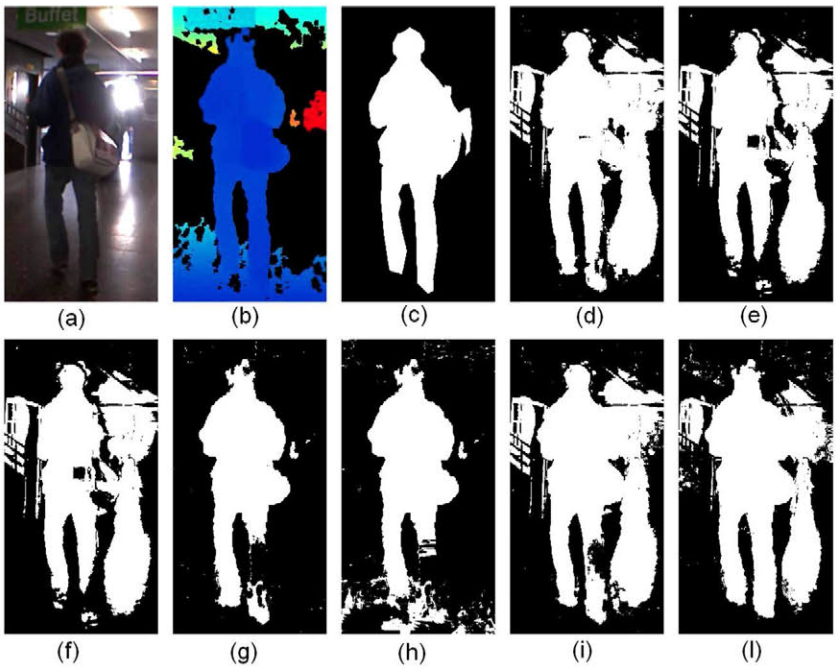

Fig. 3. Lobby2 sequence frame 195: (a) color data, (b) depth data (*nmd* pixels in black), (c) ground truth, (d) proposed $MoG - RegPRE$, (e) MoG_C, (f) MoG_{RC}, (g) MoG_D, (h) MoG_{RD}, (i) CL_W proposed by Camplani and Salgado (2013), (l) MoG_{RGB-D} proposed by Gordon et al. (1999).

Table 2
Detection accuracy obtained by analyzing the Lobby2 sequence.

	TE	FN	FP	S
$MoG - RegPRE$	**7.74**	**28.54**	**5.05**	**0.47**
MoG_C	11.26	54.74	5.63	0.31
MoG_D	7.20	56.99	0.76	0.32
MoG_{RC}	12.32	59.42	6.22	0.26
MoG_{RD}	11.21	56.01	5.42	0.25
MoG_{Bin}	8.01	67.89	0.26	0.27
MoG_{RGB-D}	8.22	17.86	6.97	0.47
$ViBe_{bin}$	7.54	56.75	1.17	0.38
CL_W	9.75	22.21	8.14	0.43

case more compact, however a higher level of false positive (e.g., the roof) is present in this case.

The results of the algorithms comparison for the sequence Lobby2 are reported in Table 2. Also in this case, the proposed algorithm allows to improve the performance obtained by the independent classifiers: by reducing the number of FN and keeping low the value of FP. The proposed algorithm allows to obtain the highest value for *S*; a similar value is obtained with MoG_{RGB-D}, however this approach leads to a higher value of TE. It is worth noting that a smaller value for TE, with respect to the proposed approach, is obtained by MoG_D and $ViBe_{bin}$, however these approaches lead to unacceptable rate of FN. It has to be underlined that also in this case the approaches based on a binary combination of foreground masks such as $ViBe_{bin}$ and MoG_{bin} lead to poor results, since an error in one of the two models affect negatively the overall performance of the system.

Lobby3 sequence, the last one of the database proposed by Spinello and Arras (2011), is presented in Fig. 4. The detected foreground obtained with the proposed method is reported in Fig. 4(d), as it can be noticed also in this case, the combination of the independent models and the region-based foreground prediction allows to efficiently fuse the depth and color information and obtain accurate foreground detections. In particular, the proposed approach improves the detection obtained separately by the color based classifiers (Fig. 4(e) and (f)) and the depth based classifiers (Fig. 4(g) and (h)). A similar result in this example is obtained also with CL_W, however a high level of FP is obtained for example, in the background wall. The algorithm MoG_{RGB} suffers the influence of the modification of the illumination conditions.

The results obtained with the different algorithms are presented in Table 3. The proposed approach guarantees for this sequence the higher value of *S* and the second lowest value for TE. As in the case of sequence Lobby2, the lowest value of TE is obtained by MoG_D, however, by considering only the depth data a high value of FN is obtained. Regarding the other state of the art algorithms, it has to be underlined that the binary-like methods (MoG_{bin} and $ViBe_{bin}$) give poor results with a low value of *S* and a high value for TE. Results similar to the one reached with the proposed approach are obtained with CL_W, as also seen in the example in Fig. 4, however this approach leads to a high level of FP. Also in this case the best performance of the proposed approach are strictly related to the region-based foreground prediction module that helps to obtain more compact foreground detections, while at the same time reducing the number of FP.

An example of one frame of the *GenSeq* sequence is shown in Fig. 5. Although in this less complex scenario the detection accuracy of the independent classifiers increases, they are still affected by different errors. Depth based classifiers (Fig. 5(g) and (h)) are characterized by noisy object boundaries and higher levels of FP. The typical errors of the color based classifiers are present also in this sequence, as it can be noticed from Fig. 5(e) and (f), where fragmented foreground regions are obtained mainly due to the color camouflage problem. The proposed approach reduces the errors that affect the independent classifiers leading to compact and refined foreground detected regions. As it can be noticed from Fig. 5(i) and (l) the algorithms CL_W and MoG_{RGB-D} are affected by a higher level of false positive.

The obtained results are showed in Table 4. The proposed algorithm achieves the lowest TE and the highest *S*, proving its superior performance. Moreover, the values of FP and FN are lower than the ones obtained using only the independent classifiers of our system. Regarding the performance of state of the art algorithms, MoG_{RGB-D}

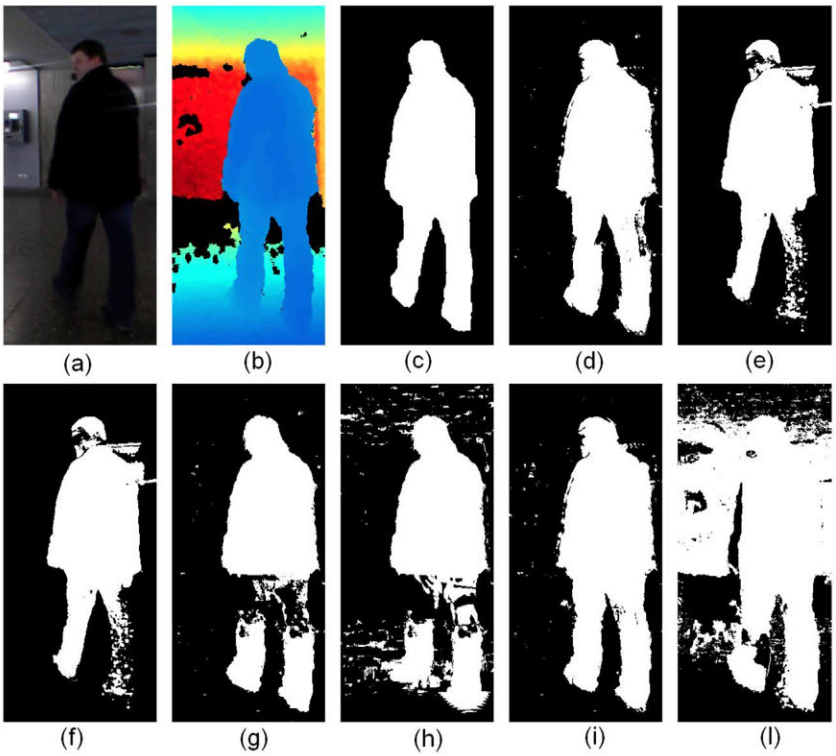

Fig. 4. Lobby3 sequence frame 420: (a) color data, (b) depth data (*nmd* pixels in black), (c) ground truth, (d) proposed $MoG - RegPRE$, (e) MoG_C, (f) MoG_{RC}, (g) MoG_D, (h) MoG_{RD}, (i) CL_W proposed by Camplani and Salgado (2013), (l) MoG_{RGB-D} proposed by Gordon et al. (1999).

Table 3
Detection accuracy obtained by analyzing the Lobby3 sequence.

	TE	FN	FP	S
$MoG - RegPRE$	**4.32**	**20.36**	**2.87**	**0.46**
MoG_C	9.19	54.74	5.06	0.21
MoG_D	3.65	31.70	1.11	0.34
MoG_{RC}	9.47	55.17	5.33	0.21
MoG_{RD}	7.08	31.21	4.89	0.23
MoG_{Bin}	11.50	58.24	7.27	0.13
MoG_{RGB-D}	48.04	9.21	51.56	0.13
$ViBe_{bin}$	25.24	41.31	23.79	0.12
CL_W	7.06	18.06	6.06	0.36

Table 4
Detection accuracy obtained by analyzing the *GenSeq* sequence.

	TE	FN	FP	S
$MoG - RegPRE$	**0.85**	**1.28**	**0.79**	**0.88**
MoG_C	2.13	8.41	1.35	0.74
MoG_D	1.61	3.70	1.35	0.81
MoG_{RC}	3.36	8.74	2.69	0.66
MoG_{RD}	2.49	3.86	2.32	0.75
MoG_{Bin}	2.03	17.01	0.16	0.74
MoG_{RGB-D}	1.93	0.63	2.09	0.79
$ViBe_{bin}$	12.39	0.65	13.85	0.44
CL_W	1.13	2.26	0.99	0.85

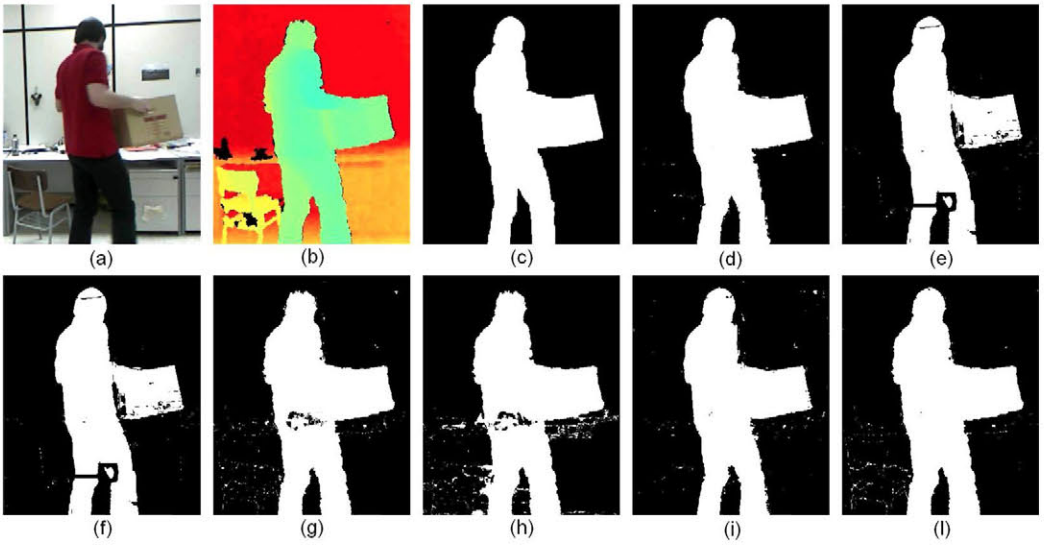

Fig. 5. *GenSeq* sequence frame 984: (a) color data, (b) depth data (*nmd* pixels in black), (c) ground truth, (d) proposed $MoG - RegPRE$, (e) MoG_C, (f) MoG_{RC}, (g) MoG_D, (h) MoG_{RD}, (i) CL_W proposed by Camplani and Salgado (2013), (l) MoG_{RGB-D} proposed by Gordon et al. (1999).

Table 5
Acronyms description.

Type	Name	Description	Type	Name	Description
Modules	*MShift*	Mean Shift	Main Equation Symbols	r	Sorting factor in MoG
	BgMod	Background Modeling		ω, σ, μ	Weight, variance and mean of the MoG distributions
	RegPreg	Depth-based Foreground Region Prediction		σ_{noise}	Variance of the distance-dependent noise
	MoE	Mixture of Expert		α	MoG learning rate
	MoG_C	Classifier based on the color per-pixel background model		DP, M, W	Decision Profile, overall support, Weight used in MOE
	MoG_D	Classifier based on the depth per-pixel background model		P_G, P_C, P_D	Edge closeness probabilities
	MoG_{RC}	Classifier based on the color region background model		t_c	Temporal Consistency Score
	MoG_{RD}	Classifier based on the depth region background model		S_h, S_{t_h}, S_t	Depth histograms sets for prediction
				b_d	Bhattacharyya distance
Data	C_t, D_t	Actual color and depth data	Algorithms in the Results Section	$MoG - RegPRE$	Proposed Approach
	$MS - C_t, MS - D_t$	Segmented depth and color data		MoG_{Bin}	Approach presented by Stormer et al. (2010)
	Bg_t	Depth background model		MoG_{RGB-D}	Approach presented by Gordon et al. (1999)
	Fg_{t-1}	Previous Foreground mask		$ViBe_{Bin}$	Approach presented by Leens et al. (2009)
	p_{fg}, p_{bg}	Predicted foreground and background probabilities		CL_W	Approach presented by Camplani and Salgado (2013)
	L	Likelihoods estimated by *BgMod*			
	nmd	non measured depth pixels			

and CL_W have comparable results with our approach. However, the proposed approach guarantees a better balance between FN and FP. As expected, the binary combination (MoG_{Bin} and $ViBe_{Bin}$) obtained with independent modules does not guarantee accurate results due to the fact that the final classification is completely compromised when one of the classifier fails.

4. Conclusions

In this paper, we present an innovative background subtraction algorithm based on the fusion of multiple region-based classifiers that processes data provided by RGB-D imagery. The proposed strategy employs different background models, based on the Mixture of Gaussian algorithm, that are built by analyzing the spatial and temporal evolution of depth and color features. These models are combined with a foreground prediction scheme that is based on a depth-histogram appearance model of the foreground regions combined with a dynamic model. The background models and the foreground prediction data are fused in a mixture of experts systems that implements a weighted average scheme that is able to improve the final detection accuracy. The main contributions of the proposed approach are therefore the following: the use of background models based on region and pixel temporal evolution, considering depth and color data, and the foreground prediction scheme based on a region-based foreground model. The results show that the proposed background subtraction approach outperforms state of the art algorithms based on RGB-D imagery. Moreover, as described with the test on databases that include difficult and elaborated scenes, the results demonstrate that it is able to efficiently detect foreground objects in very challenging situations such as strong illumination variations, RGB-D camera interferences, large region containing non-measured depth data, depth camera noise and crowded scenes.

Acknowledgements

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.

Appendix A

Table 5 summarizes the acronyms used in the paper.

References

Aach, T., Kaup, A., 1995. Bayesian algorithms for adaptive change detection in image sequences using markov random fields. Signal Processing: Image Communication 7, 147–160.

Barnich, O., Van Droogenbroeck, M., 2011. ViBe: a universal background subtraction algorithm for video sequences. IEEE Transactions on Image Processing 20, 1709–1724.

Bouwmans, T., 2011. Recent advanced statistical background modeling for foreground detection – a systematic survey. Recent Patents on Computer Science 4, 147–176.

Bouwmans, T., Baf, F.E., 2008. Background modeling using mixture of Gaussians for foreground detection-a survey. Recent Patents on Computer Science 3, 219–237.

Camplani, M., Salgado, L., 2013. Background foreground segmentation with rgb-d kinect data: an efficient combination of classifiers. Journal of Visual Communication and Image Representation, in press, http://dx.doi.org/10.1016/j.jvcir.2013.03.009.

Camplani, M., Mantecon, T., Salgado, L., 2012. Accurate depth-color scene modeling for 3d contents generation with low cost depth cameras. In: Image Processing (ICIP), 2012 19th IEEE International Conference on, pp. 1741–1744.

Camplani, M., Mantecon, T., Salgado, L., 2013. Depth-color fusion strategy for 3d scene modeling with kinect. IEEE Transaction on Cybernetics, in press, http://dx.doi.org/10.1109/TCYB.2013.2271112.

Chen, L., Wei, H., Ferryman, J., 2013. A survey of human motion analysis using depth imagery. Pattern Recognition Letters 34, 1995–2006, http://dx.doi.org/10.1016/j.patrec.2013.02.006.

Clapés, A., Reyes, M., Escalera, S., 2013. Multi-modal user identification and object recognition surveillance system. Pattern Recognition Letters 34, 799–808.

Comaniciu, D., Meer, P., 2002. Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 603–619.

Cristani, M., Farenzena, M., Bloisi, D., Murino, V., 2010. Background subtraction for automated multisensor surveillance: a comprehensive review. EURASIP Journal on Advances in Signal Processing 2010, 1–24.

Doisy, G., Jevtic, A., Lucet, E.,, Edan, Y., 2012. Adaptive person-following algorithm based on depth images and mapping. In: Proc. of the IEEE IROS Workshop on Robot Motion Planning.

Gordon, G., Darrell, T., Harville, M., Woodfill, J., 1999. Background estimation and removal based on range and color. In: IEEE Conference on Computer Vision and, Pattern Recognition, p. 464.

Han, J., Shao, L., Xu, D., Shotton, J., 2013. Enhanced computer vision with microsoft Kinect sensor: a review. IEEE Transactions on Cybernetics 43, 1318–1334, http://dx.doi.org/10.1109/TCYB.2013.2265378.

Jodoin, P.M., Mignotte, M., Konrad, J., 2007. Statistical background subtraction using spatial cues. IEEE Transactions on Circuits and Systems for Video Technology 17, 1758–1763.

KaewTraKulPong, P., Bowden, R., 2001. An improved adaptive background mixture model for real-time tracking with shadow detection. In: European Workshop on Advanced Video Based Surveillance Systems. Kluwer Academic Publishers., pp. 149–158.

Khoshelham, K., Elberink, S.O., 2012. Accuracy and resolution of Kinect depth data for indoor mapping applications. Sensors 12, 1437–1454.

Klare, B., Sarkar, S., 2009. Background subtraction in varying illuminations using an ensemble based on an enlarged feature set. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 66–73.

Kuncheva, L., 2004. Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience.

Leens, J., Barnich, O., Piérard, S., Droogenbroeck, M., Wagner, J.M., 2009. Combining color, depth, and motion for video segmentation. In: Computer Vision Systems. Volume 5815 of Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp. 104–113.

Li, L., Huang, W., Gu, I.Y.H., Tian, Q., 2004. Statistical modeling of complex backgrounds for foreground object detection. IEEE Transactions on Image Processing 13, 1459–1472.

Mahbub, U., Imtiaz, H., Roy, T., Rahman, M.S., Ahad, M.A.R., 2013. A template matching approach of one-shot-learning gesture recognition. Pattern Recognition Letters 34, 1780–1788, http://dx.doi.org/10.1016/j.patrec.2012.09.014.

Otsu, N., 1979. A threshold selection method from gray-level histograms. Systems, Man and Cybernetics, IEEE Transactions on 9, 62–66.

Polikar, R., 2006. Ensemble based systems in decision making. Circuits and Systems Magazine, IEEE 6, 21–45.

Spinello, L., Arras, K., 2011. People detection in rgb-d data. In: Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on, pp. 3838–3843.

Stauffer, C., Grimson, W., 1999. Adaptive background mixture models for real-time tracking. In: Conference on Computer Vision and Pattern Recognition. IEEE, pp. 246–252.

Stone, E., Skubic, M., 2011. Evaluation of an inexpensive depth camera for in-home gait assessment. Journal of Ambient Intelligence and Smart Environments 3, 349–361.

Stormer, A., Hofmann, M., Rigoll, G., 2010. Depth gradient based segmentation of overlapping foreground objects in range images. In: Conference on Information Fusion. IEEE, pp. 1–4.

Toyama, K., Krumm, J., Brumitt, B., Meyers, B., 1999. Wallflower: principles and practice of background maintenance. Proceedings of the Seventh IEEE International Conference on Computer Vision, 255–261.

Zivkovic, Z., van der Heijden, F., 2006. Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognition Letters 27, 773–780.