Reliable Left Luggage Detection Using Stereo Depth and Intensity Cues

Csaba Beleznai, Peter Gemeiner, and Christian Zinner
Safety & Security Department
AIT Austrian Institute of Technology, Vienna, Austria
csaba.beleznai@ait.ac.at

Abstract

Reliable and timely detection of abandoned items in public places still represents an unsolved problem for automated visual surveillance. Typical surveilled scenarios are associated with high visual ambiguity such as shadows, occlusions, illumination changes and substantial clutter consisting of a mixture of dynamic and stationary objects. Motivated by these challenges we propose a reliable left item detection approach based on the combination of intensity and depth data from a passive stereo setup. The employed in-house developed stereo system consists of low-cost sensors and it is capable to perform detection in environments of up to $10m \times 10m$ in size. The proposed algorithm is tested on a set of indoor sequences and compared to manually annotated ground truth data. Obtained results show that many failure modes of intensity-based approaches are absent and even small-sized objects such as a handbag can be reliably detected when left behind in a scene. The presented results display a very promising approach, which can robustly detect left luggage in dynamic environments at a close to real-time computational speed.

1. Introduction

Several recent incidents have indicated that unattended luggage detection poses an important security threat in transport and other critical infrastructures. Abandoned or left luggage detection represents also a key problem to surveillance personnel since accurate detection of rare abandonment events embedded into a cluttered environment is beyond the capability of a human observer.

Motivated by this challenge, the research field of visual abandoned object detection has been very active, proposing a large number of surveillance solutions employing stationary cameras as commonly encountered in public surveillance. Nevertheless, public challenges such as the PETS 2007 challenge [1] and the i-LIDS for AVSS'07 benchmark [2] focusing on left object detection have shown that common RGB sensors have difficulties, even when observations are based on multiple, partially overlapping views, to accurately detect left items at a reasonable rate of false alarms.

In this paper we propose a novel abandoned object detection approach which combines depth and intensity information to reliably detect and segment left object candidates in scenes with observation zones of up to 10 by 10 meters in size. Left item detection is a multi-stage process, where individual vision tasks involve the detection of static objects, object owners and the spatio-temporal relation between object and owner. The main scope of this work is put on the first static object detection stage.

The main contribution of the paper is given by the novel left luggage detection concept employing combined passive stereo depth and intensity cues, including algorithms capable to infer object presence from sparse depth data. Our proposed use of combined cues addresses the typical problems of the left object detection task: shadows and illumination changes, ambiguity between object drop-off and removal, occlusions, precise spatial segmentation and the presence of non-relevant stationary objects such as sitting persons. Individual cues are combined in an independent manner. An adaptive disparity-based background model is generated under consideration of non-valid pixels. Based on a comparison between octree-structured 3D data for background and current depth observations, proposals for spatial changes in the scene geometry are generated. Simultaneously to the depth-based left item hypotheses, intensity data is also used to generate stationary foreground region proposals. Individual cue based proposals are represented by single image region segments which are combined based on their spatial extent and temporal history. In order to cope with complex object types such as poorly textured or small-sized left objects, we apply a validation step in form of a weakly parametric region growing algorithm - inspired by Tian et. al [23] - exploiting intensity and depth cues. A slowly integrating gradient-based motion history is also computed to reveal non-rigid or quasi-stationary objects.

The paper is structured as follows: First, Section 2 gives a concise structured presentation on related work. Section 3 presents the proposed detection approach. The experimental setup, tested scenarios and their discussions are presented in Section 4. Finally, Section 5 concludes the paper.

2. Related work

A substantial amount of research exists on the subject of left item detection. Existing approaches can be categorized with respect to the applied detection strategy or according to the employed sensor/cue setup. These individual categories are summarized as follows:

According to detection strategies: Static object detection based on background modeling is a common technique since it is fast to compute, and background modeling concepts can be easily extended to characterize stationary foreground regions. As proposed by Tian et al. [23], in the case of Gaussian Mixture Models a certain mode of the temporally aggregated intensity distribution at a given pixel well represents static pixels. Other works rely on a temporal sequence of detected foreground regions which can be combined by accumulation [11], [17] or temporal subsampling [14] in order to derive static image regions. Such moving foreground based techniques face ambiguity problems by having no means to discriminate between object drop-off and removal. Combination of two background models computed at different frame rates or update rates, as proposed by Porikli et al. [19], better resolves this ambiguity, however, illumination variations and shadows still pose a substantial problem. A left object detection or validation strategy can be also formulated by tracking, such as in [25]. Long-term stable tracking in presence of frequent occlusions and illumination changes, however, represents a great challenge to tracking based approaches. Left object detection can be also posed as an activity recognition problem, aiming to recognize specific spatio-temporal signatures of loitering, drop-off events and subsequent static objects [6], [20], [15]. Usually such approaches require a characterization of object motion in form of low-level motion cues or tracking; however, such representations experience substantial problems with increasing object density in the scene.

According to employed sensor/cue setups: Multiple-view observation of a common calibrated ground plane introduces significant improvements [13], [7], [10]. Low-level ambiguities such as occlusions, illumination variations and shadows can be better resolved, while detection and tracking methodologies gain in robustness, thus contributing to a more reliable characterization of the spatio-temporal scene context required to recognize static objects and the status of abandonment. Another way to approach the left object detection task is to assess changes in the scene depth. Recent sensor technology developments, such as Time-of-Flight cameras and the Microsoft Kinect [3], have promoted the use of depth sensing sensors for scene analysis. These sensors employ active illumination enabling the computation of high-quality dense depth maps. Mainly due to the limited spatial range and resolution, and the problems of active sensing in presence of sunlight, such sensors have been primarily applied to the indoor human detection task [22], [9]. There is little work on left object detection [18] by means of depth sensing sensors.

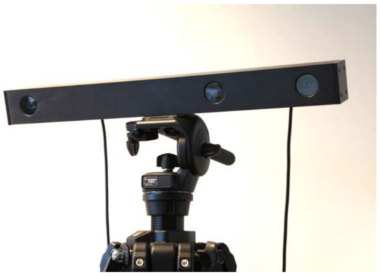

Figure 1. Our customized trinocular stereo camera setup with a baseline of $40cm$ (between the left and right cameras). A commercially available alternative (no affiliation) can be found at [5].

In light of the existing state of the art it is apparent that multi-view and depth sensing overcomes many of the commonly encountered difficulties of monocular vision based techniques. Passive stereo sensors bring additional advantages in terms of large observable area, high spatial resolution and the capability of outdoor operation under sunlit conditions.

3. Proposed approach

In this section first we briefly describe the visual input our proposed approach requires. Next, we give a overall view on the employed concepts. Finally we describe the individual algorithmic components in detail.

Image data: We use an in-house developed sensor (Figure 1) to extract intensity information, employing a canonical stereo setup (three monochrome cameras mounted in parallel), with a baseline of $0.4m$ between the two cameras located at the ends of the rig. The board-level industrial cameras have a USB2 interface and the resolution of the sensor is 1280×1024 pixels, resampled to 1150×920 with 8 bit quantization. This trinocular camera setup is calibrated offline. The stereo matching process outputs depth data alongside with rectified intensity images, congruent to the depth image. Depth information is computed via a pyramidal implementation of a Census-based stereo matching algorithm, which is an explicit adaption and optimization of the well-known Census transform in respect to embedded real-time systems in software. Depth is computed for all three available baselines thus improving the quality of obtained depth map at the different spatial ranges. At the given resolution, the sensor delivers approximately 10*fps*, when stereo computation is performed on a modern PC.

Overview: Our detection method independently computes depth and intensity based left object proposals, which are combined and validated in a later processing stage. Figure 2 depicts the simplified overview of our algorithmic flow. The main motivation for the employed independent processing scheme is the following: intensity data is de-

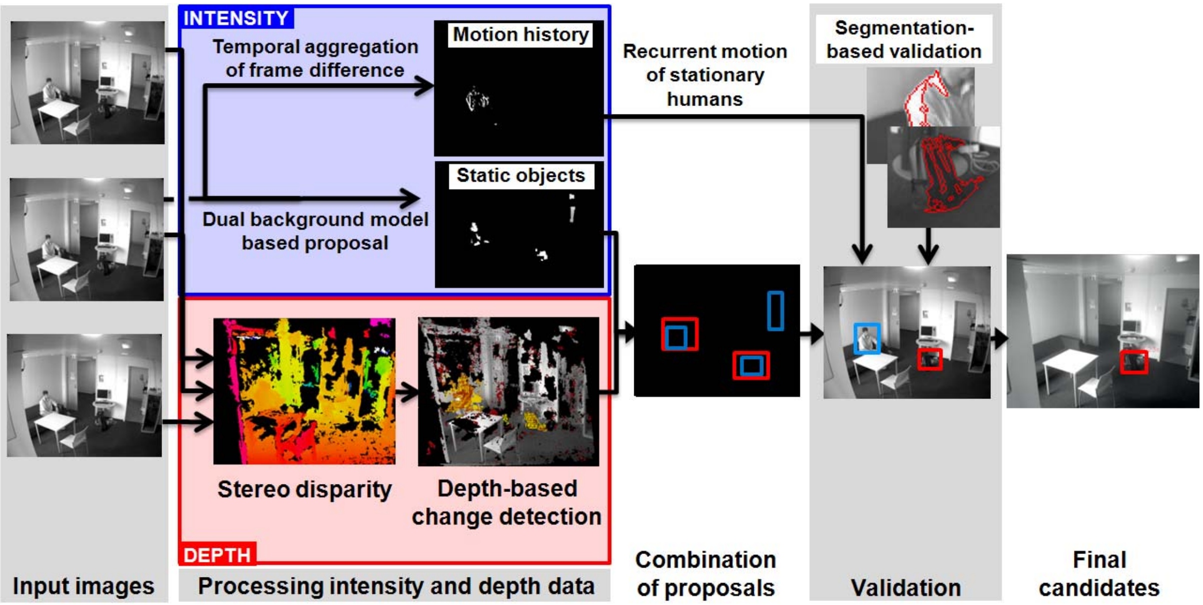
INTENSITY
Temporal aggregation of frame difference
Motion history
Static objects
Dual background model based proposal
Recurrent motion of stationary humans
Segmentation-based validation
Stereo disparity
Depth-based change detection
DEPTH
Input images
Processing intensity and depth data
Combination of proposals
Validation
Final candidates

Figure 2. Overview on the algorithmic concept. Mutually agreeing intensity- (blue) and depth-based (red) proposals are combined and validated. Proposals not passing the validation step are eliminated (cyan) in the validation stage.

fined everywhere in an acquired image, however, it exhibits a high degree of uncertainty by being sensitive to photometric variations. Computed depth, on the other hand, is available only at image regions possessing sufficient structure or texture, while accurately characterizing the scene geometry. Depth is not completely decorrelated from photometric variations, since in underexposed or overexposed image regions no depth can be estimated. By combining left object proposals from the different cues we strive to detect agreeing proposals and thus to minimize the ambiguity with respect to their source of origin, whether they are true left objects causing local depth and intensity changes or something else. A subsequent validation step also pursues this objective by performing spatial (local segmentations) and temporal (motion integration) characterization of left object proposals.

Our framework exhibits following constraints:

- requires a calibrated, stationary stereo camera, and
- the size of detectable left objects is limited by the spatial resolution of the stereo camera. The currently employed resolution allows for the detection of a small-sized backpack up to 10 meters from the camera.

As shown in Figure 2, first the disparity and intensity data from the stereo camera are used to compute background models. The current and background model disparities are converted into depth data and represented as individual octree-based voxel grids, which are used to detect and segment spatial changes. Spatial changes are accumulated in time, where long term deviations are marked as left object proposals. The intensity data is used within a dual background model [19] to generate left object proposals. Mutually agreeing depth- and intensity-based proposals are combined based on a simple region overlap criterion (see Figure 2). In a last validation step remaining proposals are examined with respect to motion patterns caused by quasi-stationary objects (e.g. humans) and plausible, ”object-like” segmentations. The segmentation step should discriminate from faulty proposals caused by highlights or underexposed areas with no structure. Validated proposals persisting longer than a predefined duration are labeled as detected left objects.

Next, individual algorithmic parts are described in detail.

3.1. Background modeling and change detection

Our disparity-based background modeling scheme employs a simple running average technique [24] which also considers the validity of the disparity value at a given pixel. A slow adaptation rate ensures that dropped-off and displaced objects only slowly become part of the reference disparity background. The disparity-based background model and the current disparity frame are converted to depth values according to these equations:

$$\begin{aligned} Z &= f * B \,/\, d; && (1)\\ X &= (u - px) * Z \,/\, f; && (2)\\ Y &= (v - py) * Z \,/\, f; && (3) \end{aligned}$$

In the above equations f is the focal length, B is baseline and d is the value of the disparity. u and v are the pixel locations in 2D image. px and py are the coordinates of the principal point. X, Y and Z are the output coordinates of

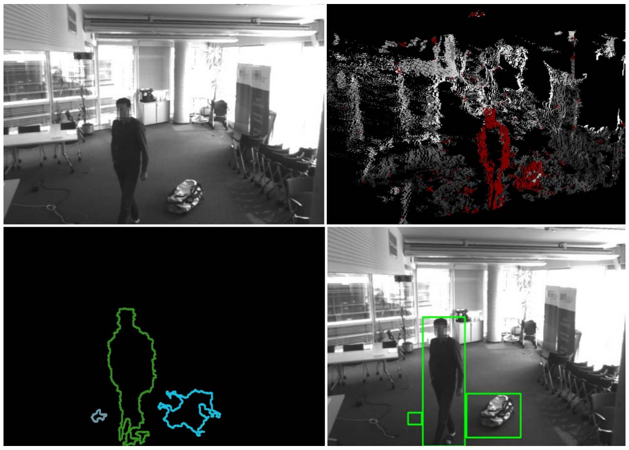

Figure 3. Change detection in depth data: Intensity image (top left); Momentary (not aggregated) changes (red) with respect to a reference depth (top right); Obtained segmentation in the 2D image space (bottom left); Bounding box representation (bottom right)

3D point. The intensity value at the coordinate u and v is taken from the rectified image.

At the same time an intensity-based background model is built using the method of Zivkovic et al. [26], [4]. The depth and intensity values for background and the current frame are used to derive joint depth-intensity representations in form of an octree [12]. The reference octree representing the background model of the scene and the octree computed for the current frame are compared, where both octrees can differ in size, resolution, density and point ordering. By recursively comparing the octree-based tree structures, spatial changes represented by differences in the voxel configurations are identified. An example of the computed depth change is displayed in Figure 3 along with its segmentation result.

An additional Motion History cue is computed to represent motion occurring at a large time scale. The measure is used later in the validation phase to assess the quantity long-term integration motion within a given rectangular image region. Inspired by the Motion History concept [8], [17], we compute the running average of inter-frame differences of the norm of image gradients. A slow integration time ensures that transient objects do not accumulate a large response and quasi-stationary objects (people standing, sitting, moving vegetation) on the other hand produce a marked Motion History signature.

3.2. Depth-based proposal generation

Upon detecting depth-based scene changes, it is necessary to segment and label candidate objects. Labeling of the candidate objects can be performed either in depth or in intensity data. Due to computational constraints, labeling in our framework is performed in the 2D image space. Every point in the depth data exhibiting a change is reprojected into the image space. Due to this reprojection step temporal aggregation and spatial segmentation become easy and computationally less demanding. Temporal aggregation is performed accordingly by creating an image accumulator where momentary changes are inserted and associated over time. If a spatial change is visible in the view of the camera, then the accumulator is compared to the newly observed changes and upon correspondence respective entries are incremented. In case of occlusions or object removal the accumulator is decremented.

A temporal association between registered depth-based changes aims at detecting strongly correlating stationary candidates and distinguishing them from transient phenomena. To represent individual candidates, both in the accumulator and in the current change detection map, following two measures are employed:

- center of the fitted rectangle and
- ratio of area R.

The area ratio R is equal

$$R = Area_S / Area_N, \tag{4}$$

where $Area_S$ is the area of a stored and $Area_N$ of a new candidate. In case of a relative deviation below a threshold T the accumulator entry and the new candidate are associated and the corresponding accumulator entries are incremented by a unit value.

In order to support proposal generation by depth information, a volume estimation step is applied if an accumulator entry reaches a number of predefined observations N (N = 5 in our case). Volume measurement is based on the estimation of the convex hull (see Eq. 5) from the corresponding depth values:

$$\left\{ \sum_{i=1}^{|S|} \alpha_i x_i \;\middle|\; (\forall i : \alpha_i \geq 0) \wedge \sum_{i=1}^{|S|} \alpha_i = 1 \right\}, \tag{5}$$

where this convex hull of a finite point set $S \in R^n$ forms a convex polytope. Each point x_i in S is assigned a weight or coefficient α_i in such a way that the coefficients are all non-negative and sum to one, and these weights are used to compute a weighted average of the points [21].

The estimate of an approximated volume for a candidate region in the accumulator enables the framework to discard too large or too small objects. In the present framework such a filtering is employed only for very small-sized objects, there is no upper limit set for object size. Generated regions in the accumulator image represent the depth based proposals (red regions in Figure 2).

3.3. Intensity-based proposal generation

Using the previously (Section 3.1) described background modeling scheme [26] we generate intensity-based candidates by the dual background model of Porikli et al. [19]. Intensity-based proposals represent a sensitive way to detect stationary candidates, implying that even low-contrast and small-sized objects are detected; however, at the expense of detecting many false alarms. This high sensitivity, nevertheless, results in a high recall which is essential for fusing intensity- and depth-based proposals using mutual agreement.

3.4. Combination of proposals

Given the two sets of rectangles generated by depth and intensity-based detection, we compute the following pairwise overlap ratios:

$$r = \frac{area(B_d \cap B_i)}{area(B_d \cup B_i)}, \tag{6}$$

where B_d and B_i are two overlapping bounding boxes generated from depth and intensity data, respectively. $B_d \cap B_i$ and $B_d \cup B_i$ represent the intersection and the union of the two bounding boxes, respectively. If the overlap ratio r exceeds 50%, a match is declared. A one-to-one match is not enforced, thus several proposals of one cue can be matched to a bounding box of the other cue.

3.5. Validation of proposals

In the validation step we apply two validation mechanisms. A recurrent motion based validation step uses the aggregated Motion History information (Section 3.1). For each proposal region an area-normalized Motion History mean is computed. Proposal regions exceeding an experimentally determined threshold T_{mh} are discarded.

A segmentation-based validation step is introduced in order to detect proposals lacking any boundary, structure or texture. Multiple region growing algorithms are started from several seed point locations defined at K-by-K grid points (K = 3) within the proposal rectangle and applied to the intensity image. Since the stopping criterion of region growing is a sensitive parameter, we employ the criterion of Maximum Stability [16] to find consistent segments. By incrementally performing segmentation with increasing stopping criterion (similarity threshold) within a predefined range and using the step Δ, we compute the stability value S:

$$S(R_i^g) = (|R_j^{g-\Delta}| - |R_k^{g+\Delta}|) \ / \ |R_i^g| \tag{7}$$

$|.|$ denotes the cardinality (area) of a region. R_i^g is a region obtained with a similarity threshold g, while $R_j^{g-\Delta}$ and $R_k^{g+\Delta}$ are regions with decreased and increased similarity thresholds, respectively. A local minimum of the stability value S during region growing with increasing similarities implies a maximally stable region. In our validation step we retain the first found local minimum (if any) and verify whether it is contained within the proposals rectangle. If one of the multiple segmentation attempts (starting from the grid) meets this criterion, the proposal is accepted as a valid candidate.

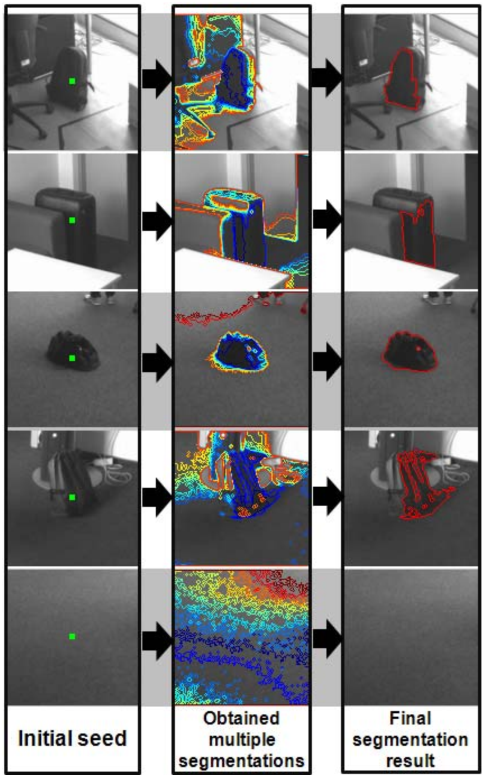

Figure 4. Example segmentation results obtained by the used weakly-parametric region growing validation step. Shown: Initial seed points (left), color-coded multiple segments according to the varying similarity threshold (center) and the obtained maximally stable segment (right).

Figure 4 shows various segmentation results for a single seed point. As it can be seen, for objects certain parts (facets, patches) are found as stable segments, while in a weakly structured area (last row of Figure 4) the growth of incrementally computed segments does not slow down due to absence of object-delineating boundaries.

4. Experimental results and discussion

In this section, we present the experimental setup and illustrate and discuss the detection results.

Dataset: To our best knowledge there is no publically available dataset for depth-based left object detection. For this reason we recorded 6 indoor scenarios depicting several events of object (luggage) drop-off in an office and lab environment. Unfortunately, we could not reproduce the complexity of a public location (crowd, clutter), nevertheless,

Complexity aspects	DOOR	MEETING	COFFEE TABLE	CORRIDOR	LAB	TWO DOORS
Illumination (low light, saturation, changes)	×		×			×
Dynamic occlusions		×		×		
Non-relevant static object (person, moved door, chair)		×	×		×	×
Small-sized left object				×		
Number of frames	1995	12615	2437	2643	2947	2460
Number of annotated static regions	1	6	4	2	5	4
Number of true left objects	1	5	3	2	3	2

Table 1. Parameters of the video sequences used for the evaluation.

Performance measure	DOOR	MEETING	COFFEE TABLE	CORRIDOR	LAB	TWO DOORS
Precision (P)	1	0.71	0.75	1	0.6	0.5
Recall (R)	1	1	1	1	1	1

Table 2. Obtained quantitative results.

we tried to introduce several complicating factors, which are well-known to cause failures in published left item detection concepts. The dataset was recorded by our vision sensor (Section 3). Table 1 summarizes the complexity aspects contained in the individual datasets and provides information on the associated annotation.

Results: In all of our experiments we set a time window of 25s which is required for a validated proposal to raise a left item alarm. Qualitative results obtained for the six test sequences are shown in Figure 5. Final detection results are depicted in the last column of Figure 5. Obtained results imply that depth as a cue for the detection task is very valuable. Depth-based detection was consistent in all of the sequences and showed only occasional local instabilities in presence of highlights. For example, appearance (opening door) and disappearance (closing door) of highlights in the TWO DOORS sequence generates a sudden appearance of valid disparity pixels, which are interpreted as a change in the scene geometry. Most of these changes are eliminated in the validation phase, however, complex cases can be easily imagined (e.g. background with tiles) where our validation mechanism would fail. The intensity cues are important for low-contrast and small-sized objects (DOOR, CORRIDOR), where the depth cue often results in oversegmented region proposals. Dynamic occlusions (MEETING) are handled well, although an explicit depth-based analysis of proposals or depth-ordering has not been employed. A very complex group of situations is given by transferred objects, such as a pushed rolling chair or doors opening and closing. Such objects show up as static objects, and further information would be needed to recognize them as non-relevant. Keeping track of all scene objects over time is one possibility, but it requires a detailed scene segmentation and analysis. Measured geometric attributes such as size, height or compactness might also represent a viable solution, but still do not fully discriminate from luggage items.

The validation mechanisms well rejected quasi-stationary object motion and objects without a boundary or texture. Nevertheless, in the presence of large amounts of motion clutter unseen failure modes might arise.

By manually annotating the test datasets we performed a quantitative comparison of detection results. Using a bounding box representation for ground truth and detection results, we employed the bounding box overlap criterion (analogously to Equation 6) to assess the detection performance. The detection performance is gauged by means of Precision (P) and Recall (R). Precision is referred to how many returned left items are relevant and equal to $\frac{tp}{tp+fp}$. Recall is referred to what fraction of relevant left items was found and equal to $\frac{tp}{tp+fn}$. tp, fp and fn are the true, false and missed detections, respectively. Table 2 displays the obtained quantitative results for all datasets. As it can be seen, a high recall is achieved, however, the previously mentioned failure modes of transferred objects and occasional highlights generate few false alarms, which reduce the obtained precision scores.

Short term occlusions are handled well, but the evidence accumulation in a backprojected image space has disadvantages when occlusions last longer. If a proposed left item is occluded by a dynamic occluder, using the current approach its importance will decrease and after a while the candidate might disappear. An accumulation of left item proposals in the 3D space would represent an improved concept, where occlusions can be detected and proposals can be handled accordingly.

The proposed framework runs at 5 fps on a modern PC. 10 fps is reached when computing only the stereo disparities. The accomplished 5 fps is sufficient to gather a large number of depth and intensity based visual information over

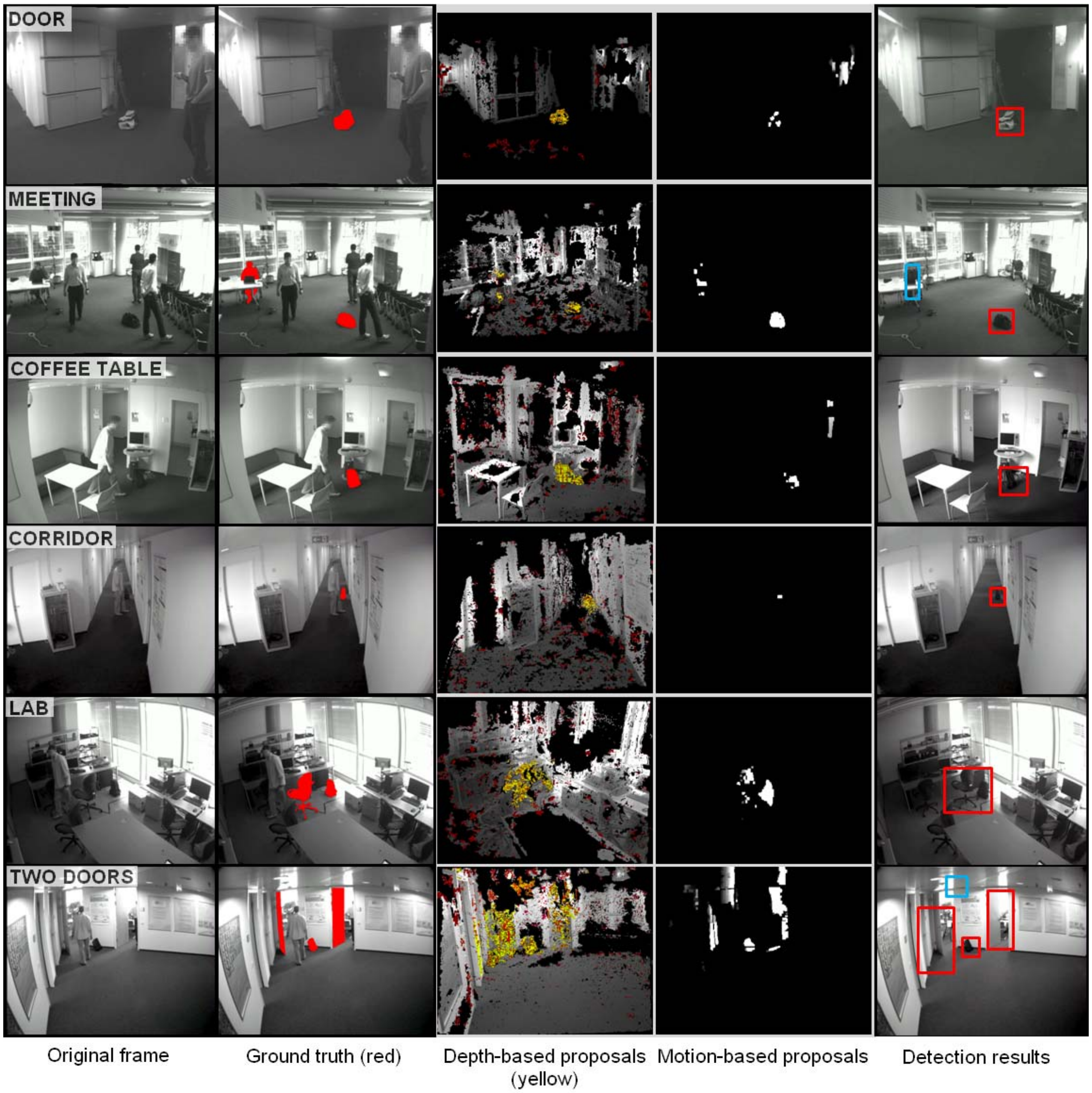

Figure 5. Qualitative detection results.

the duration of the time window necessary for generating a decision.

5. Conclusion

In this paper, we propose a novel left luggage detection framework. The framework uses intensity information and disparities from a stereo camera setup. The presented combination of intensity- and depth-based cues exhibits promising performance on a limited set of data. The accomplished near real-time run-time performance allows for practically relevant deployment in indoor and outdoor scenes with observation zones up to 10m from the camera.

Future improvements will target a tighter integration of depth and intensity cues in form of 3D aggregation of evidence and detailed occlusion analysis.

6. Acknowledgements

Part of the research leading to these results has received funding from the European Unions Seventh Framework Programme managed by REA (Research Executive Agency) (FP7/2007-2013) under grant agreement no. FP7-SME-2012 - DSenS.

References

[1] http://www.cvg.rdg.ac.uk/PETS2007/. 1

[2] http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html. 1

[3] Microsoft Corp. Kinect for XBOX. http://www.xbox.com/en-us/kinect. 2

[4] Background Subtraction Library: http://code.google.com/p/bgslibrary/. 4

[5] Point Grey Research stereo 3d cameras: http://ww2.ptgrey.com/stereo-vision/bumblebee-xb3. 2

[6] H. Ardö and K. Åström. Multi sensor loitering detection using online viterbi. In *Proc. IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance*, 2007. 2

[7] E. Auvinet, E. Grossmann, C. Rougier, M. Dahmane, and J. Meunier. Left-luggage detection using homographies and simple heuristics. In *Proc. IEEE Workshop on Performance Evaluation in Surveillance and Tracking*, 2006. 2

[8] A. F. Bobick, J. W. Davis, I. C. Society, and I. C. Society. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:257–267, 2001. 4

[9] W. Choi, C. Pantofaru, and S. Savarese. Detecting and tracking people using an rgb-d camera via multiple detector fusion. In *ICCV Workshops*, pages 1076–1083. IEEE, 2011. 2

[10] J. Ferryman, D. Hogg, J. Sochman, A. Behera, J. A. Rodriguez-Serrano, S. Worgan, L. Li, V. Leung, M. Evans, P. Cornic, S. Herbin, S. Schlenger, and M. Dose. Robust abandoned object detection integrating wide area visual surveillance and social context. *Pattern Recogn. Lett.*, 34(7):789–798, May 2013. 2

[11] S. Guler and K. Farrow. Abandoned object detection in crowded places. In *Proc. IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance*, Proc. IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance, pages 99–106, 2006. 2

[12] A. Knoll, I. Wald, S. Parker, and C. Hansen. Interactive isosurface ray tracing of large octree volumes. In *Proceedings of the 2006 IEEE Symposium on Interactive Ray Tracing*, pages 115–124, 2006. 4

[13] N. Krahnstoever, P. Tu, T. Sebastian, A. Perera, and R. Collins. Multi-view detection and tracking of travelers and luggage in mass transit environments. In *Proc. IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance*, 06 2006. 2

[14] H.-H. Liao, J.-Y. Chang, and L.-G. Chen. A localized approach to abandoned luggage detection with foreground-mask sampling. In *AVSS*, pages 132–139. IEEE Computer Society, 2008. 2

[15] S. Lu, J. Zhang, and D. D. Feng. Detecting unattended packages through human activity recognition and object association. *Pattern Recognition*, 40(8):2173 – 2184, 2007. Part Special Issue on Visual Information Processing. 2

[16] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference*, pages 36.1–36.10. BMVA Press, 2002. 5

[17] D. Ortego and J. C. SanMiguel. Stationary foreground detection for video-surveillance based on foreground and motion history images. In *Proceedings of the 2013 IEEE International Conference on Advanced Video and Signal Based Surveillance*, AVSS '13, Washington, DC, USA, 2013. IEEE Computer Society. 2, 4

[18] J. Owens. Object detection using the Kinect. *Technical Note, U.S. Army Research Laboratory*, 2012. 2

[19] F. Porikli, Y. Ivanov, and T. Haga. Robust abandoned object detection using dual foregrounds. *EURASIP J. Adv. Signal Process*, 2008, Jan. 2008. 2, 3, 5

[20] P. C. Ribeiro, P. Moreno, and J. Santos-Victor. Detecting luggage related behaviors using a new temporal boost algorithm. In *Proc. IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance*, Proc. IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance, 2007. 2

[21] R. B. Rusu and S. Cousins. 3D is here: Point Cloud Library (PCL). In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011. 4

[22] L. Spinello and K. O. Arras. People detection in rgb-d data. In *Proc. of The International Conference on Intelligent Robots and Systems (IROS)*, 2011. 2

[23] Y. Tian, R. S. Feris, H. Liu, A. Hampapur, and M.-T. Sun. Robust detection of abandoned and removed objects in complex surveillance videos. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 41(5):565–576, 2011. 1, 2

[24] F. Tombari, S. Mattocia, L. di Stefano, and F. Tonelli. Detecting motion by means of 2d and 3d information. In *Proceedings of ACCV'07 Workshop on Multi-dimensional and Multi-view Image Processing*, volume 1, pages 10 – 176. Asian Federation of Computer Vision, November 19, 2007. 3

[25] P. L. Venetianer, Z. Zhang, W. Yin, and A. J. Lipton. Stationary target detection using the objectvideo surveillance system. *Advanced Video and Signal Based Surveillance, IEEE Conference on*, 0:242–247, 2007. 2

[26] Z. Zivkovic and F. van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recogn. Lett.*, 27(7):773–780, May 2006. 4, 5