TEC2011-25995 EventVideo (2012-2014)

Strategies for Object Segmentation, Detection and Tracking in Complex Environments for Event Detection in Video Surveillance and Monitoring

D5.3

DESCRIPTION OF EXPLORATION OF THE USE OF KINECT'S DEPTH DATA IN ANALYSIS STAGES

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by

AUTHOR LIST

Marcos Escudero-Viñolo	marcos.escudero@uam.es
Juan C. SanMiguel	Juancarlos.sanmiguel@uam.es

CHANGE LOG

Version	Date	Editor	Description
0.1	24-10-2014	Marcos Escudero-Viñolo	2.2, 3 (partially)
0.2	12-11-2014	*Juan C. SanMiguel*	2.1
0.3	28-11-2014	Marcos Escudero-Viñolo	1
0.4	11-12-2015	*Juan C. SanMiguel*	3
1.0	12-12-2014	*José M. Martínez*	First version

CONTENTS

1. Introduction

This document describes the systems derived from a feasibility-study carried out in the VPU. This study is focused in the use of Kinect-data as complementary information for two of the analysis stages defined in the scope of this project: detection of stolen and abandoned objects and object recognition.

1.1. Document structure

This document is composed of the following chapters:

Chapter 1: Introduction to this document.

Chapter 2: Contributions developed.

Chapter 3: Conclusions and future work.

2. Contributions

This chapter compiles the contributions developed in the VPU in the scope of this project.

2.1. Depth-based detection of abandoned and stolen objects in indoor scenarios

2.1.1. Motivation

Current systems for detecting abandoned or stolen objects in video-surveillance are often based on the analysis of color images (visible spectrum). Most of the proposed approaches aim to detect stationary foreground over time by using a background model. Hence, background subtraction (BS) techniques are widely used in these systems as pivotal step for the additional analysis performed to detect the object of interest. However, BS suffers from many problems such as camouflage, noise, illumination changes,… presenting a major hurdle for the success of these systems in accurately locate the suspicious objects in time and space (see Figure 1).

This work contributes to the state-of-the-art in abandoned and stolen object detection by using depth information as an alternative source to identify the moving and static objects of the scene. Note that depth information does not have the same problems as previously presented and can be easily extracted with cheap sensors such as the Microsoft Kinect.

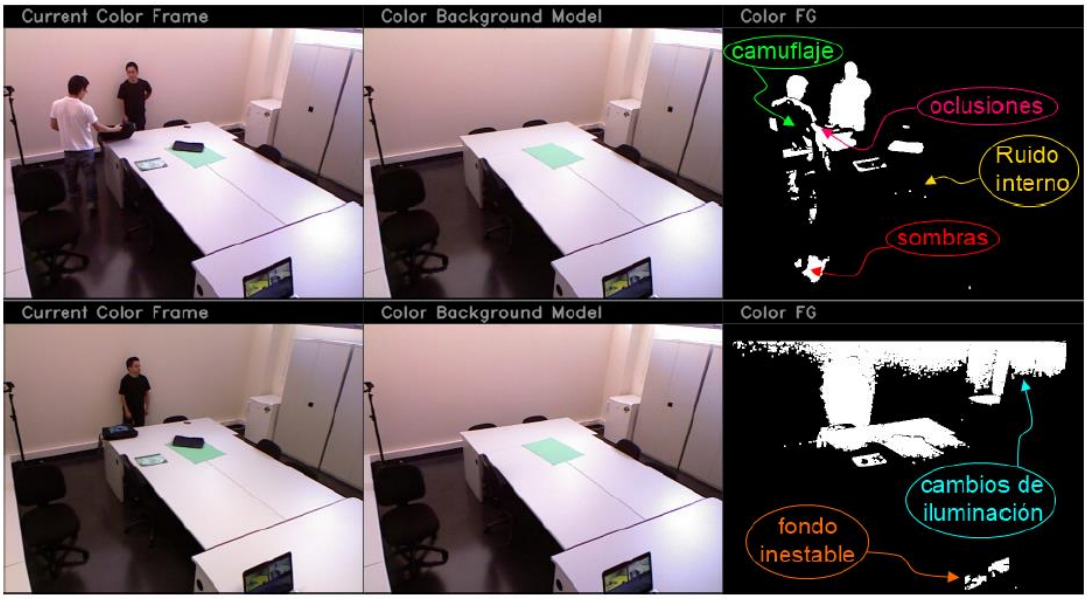


Figure 1. Common problems in systems for abandoned object detection based on background subtraction.

2.1.2. Approach description

A prototype is designed for the abandoned and stolen object detection task (see Figure 2). Once the frames are acquired, a foreground segmentation stage is performed [1]. Then, this foreground is filtered by removing shadows and highlights using the HSV colour space [2] and by removing salt&pepper noise[3]. Then, the stationary foreground is detected over time by an analysis of the foreground persistence and the existing motion via subsampling techniques as in

[4]. The result is a binary mask with stationary regions that correspond to potential objects of interest. Finally, such regions are discriminated between abandoned and stolen by studying the contrast among the boundaries of the detected regions [5].

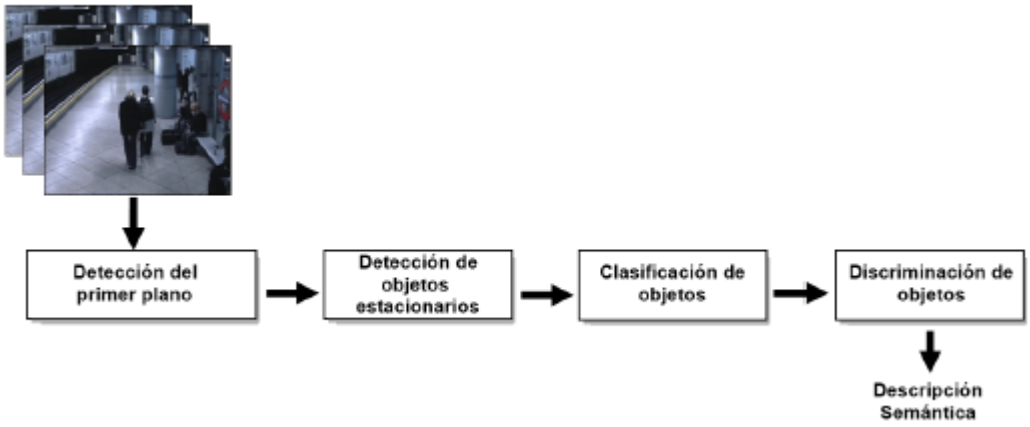


Figure 2. Initial prototype designed for abandoned and stolen object detection.

Up until now, the previously described system uses colour information. The depth information via the Kinect sensors is integrated in this system to enhance its capabilities and overcome the visual challenges that affect the BS stage.

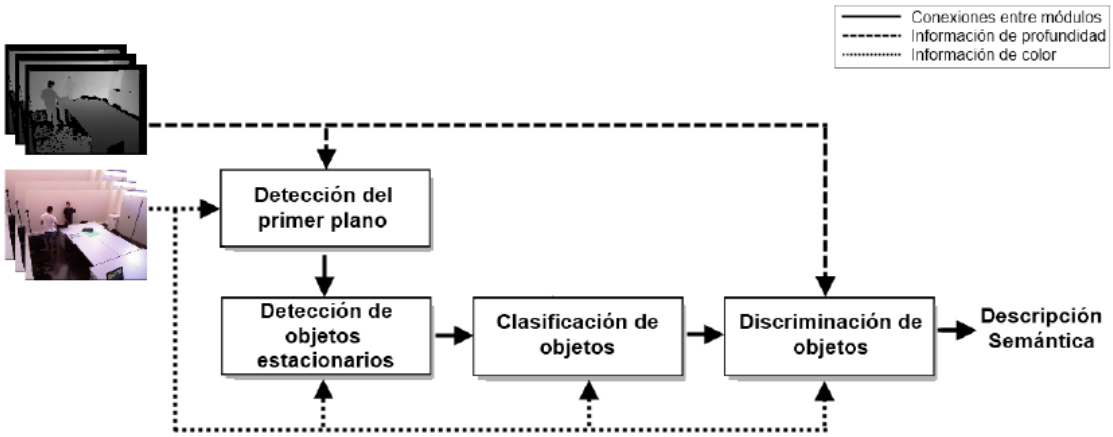


Figure 3. Modified prototype for depth-based abandoned and stolen object detection

By using the Kinect sensor, the goal is to increase the robustness of the system for abandoned and stolen object detection. Due to the use of the Kinect, the resulting system only operates for indoor scenarios as depth information cannot be accurately captured in outdoor settings. Once the colour and depth frames, we use such information for detecting the potential objects of interest and for discriminating among abandoned and stolen.

2.1.2.1. Depth-based foreground segmentation

A depth-based BS approach is proposed in the context of this work. We model each pixel depth by Gaussian described by its mean and variance over a set of training values. However, the Kinect sensor may provide an empty depth for certain pixels (see Figure 4). Therefore, this problem complicates the design of a BS approach as background pixels may contain wrong values when the sensor reports zero and non-zero depth values for the same pixel. We address

this problem by filtering out all zero-depth values. Additionally, we compute a flickering mask that indicates which pixels have zero and non-zero depth values. This mask is later used for detecting the foreground based on depth information.

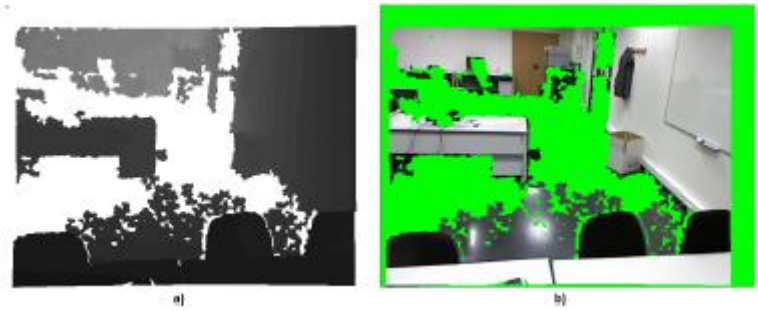


Figure 4. Depth image captured in a specific instant a) Depth b) Colour image. Green and White colours indicate where depth is not retrieved by the Kinect sensor.

After detecting the foreground using colour and depth information (both based on Gaussian models) as depicted in Figure 5, we proceed to combine them and obtain a final foreground detection that is used by the system.

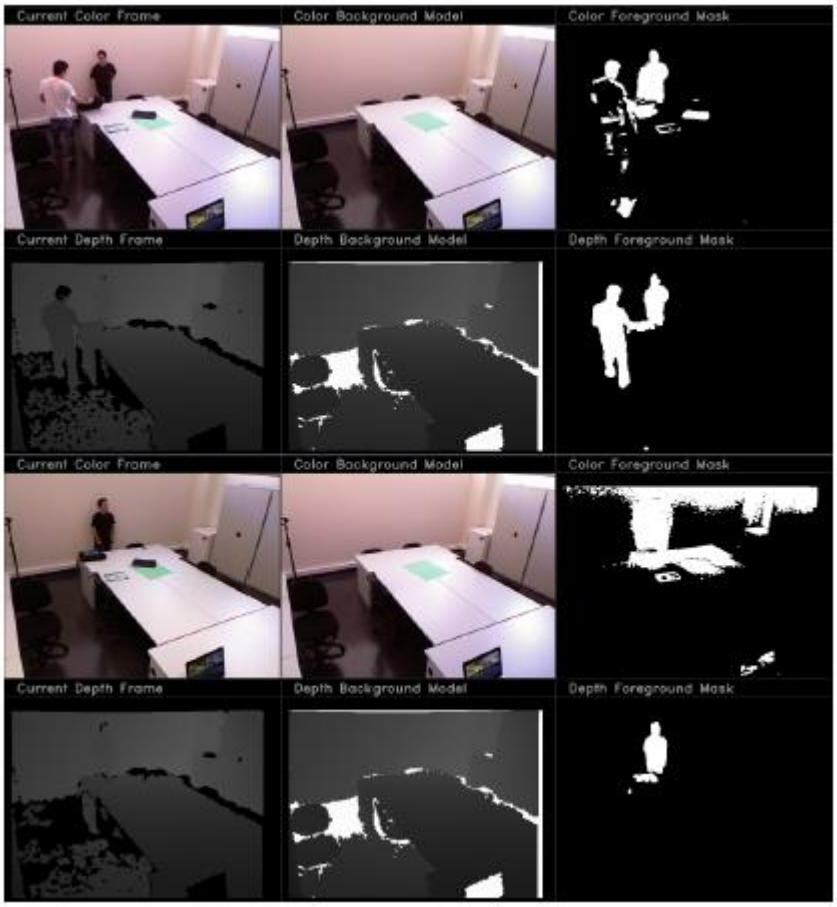


Figure 5. Examples for foreground masks using depth and colour information.

For the combination, we first copy the regions (blobs) of the colour-based foreground to the depth-based one. Thus, we include the regions with low depth discrimination with respect to its neighbourhood that usually correspond to small regions in the image (see Figure 6). Note that

directly including all colour blobs in the depth ones will be presenting similar problems to the abovementioned ones. We assume that large regions such as people or big objects, they will always appear in the depth-based mask.

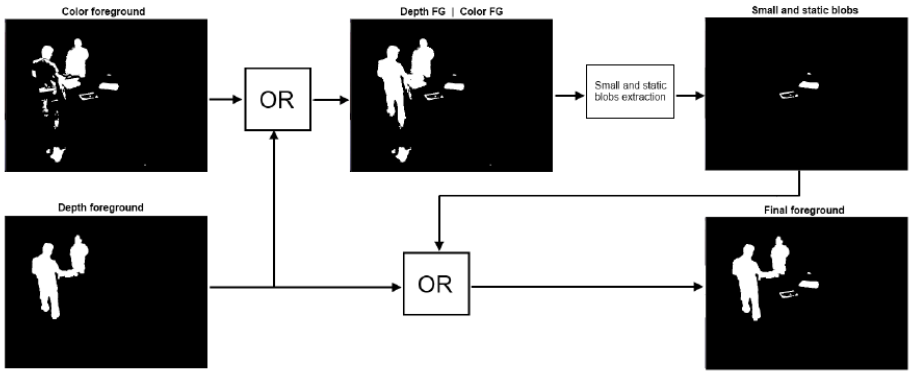


Figure 6. Combination of depth-based and colour-based foreground masks.

2.1.2.2. Depth-based discrimination

Finally, the second use of depth information is for the discrimination where the boundary energy is examined similarly as done for colour in the initial prototype. As basic rule is defined over the pixel variability as defined in Figure 7. An example is given in Figure 8.

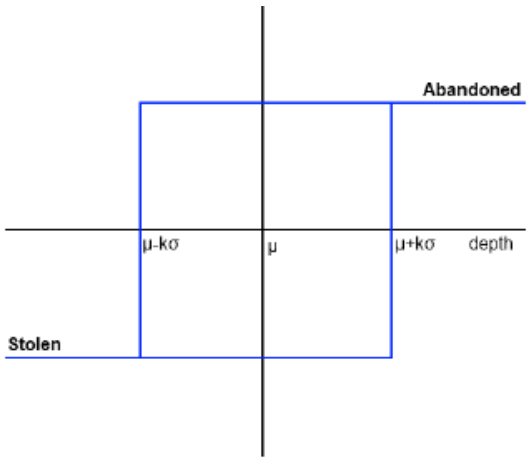


Figure 7. Rule for depth-based discrimination between abandoned and stolen.

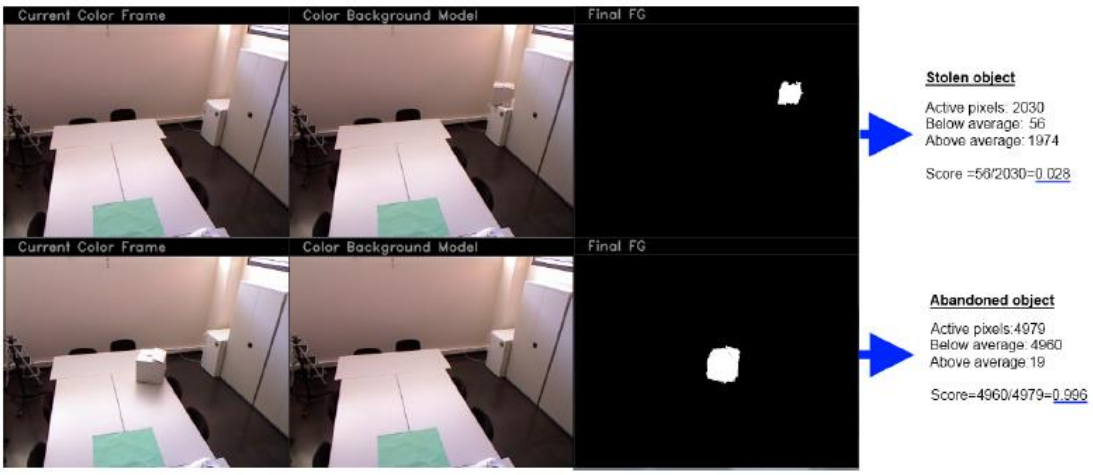


Figure 8. Example for depth–based discrimination between abandoned and stolen.

2.1.3. Experimental validation

A dataset is compiled to evaluate the proposed depth-based system and compare its capabilities against the traditional colour-based system. The following figure describes the dataset.

Secuencia	Nº abandonos	Nº robos	Nº *Frames*	Descripción	Problemas en color	Problemas en profundidad
stolen_box.oni	0	1	914	Robo de una caja	Camuflaje y sombras	Ninguno
abandoned_box.oni	1	0	776	Abandono de una caja	Camuflaje y sombras	Ninguno
stolen_bin.oni	0	1	1120	Robo de una papelera	Camuflaje y sombras	Camuflaje
abandoned_bin.oni	1	0	767	Abandono de una papelera	Camuflaje y sombras	Camuflaje
meeting_room.oni	4	0	1407	Abandono de varios objetos (carpeta, maletín, bolso de portátil y teléfono móvil)	Camuflaje, sombras, oclusiones y cambios de iluminación	Camuflaje
Illumination_change.oni	1	0	550	Abandono de una pila de libros en un entorno con iluminación variable.	Camuflaje y cambios de iluminación	Ninguno
interaction.oni	0	0	2658	Interacción entre dos personas	Camuflaje, sombras, oclusiones y cambios de iluminación	Camuflaje y oclusiones
entrance.oni	0	0	718	Personas accediendo y saliendo de un recinto	Camuflaje, sombras, oclusiones y cambios de iluminación	Oclusiones
sunshine.oni	0	0	10113	Salón interior con fuerte iluminación solar	Cambio de iluminación (solar)	Cambio de iluminación (solar)

Figure 9. Overview of the generated dataset for depth-based abandoned and stolen object detection.

We used Precision and Recall metrics for assess the performance of the compared systems. Both metrics apply to two tasks (foreground detection and abandoned/stolen detection) which are described as follows.

2.1.3.1. Depth-based foreground segmentation

We compare the foreground segmentation using only colour, only depth and the proposed combination of both. The following Table summarizes the results of the experiments for the generated dataset. It can be seen that on average, a noticeable improvement is achieved by the proposed combination. Moreover, note that only using depth-data (second column), it improves the performance of colour-based foreground segmentation by overcoming illumination changes and camouflage effects which frequently occur in the dataset.

Secuencia	Parámetro	Color	Profundidad	Combinación
meeting_room.oni	**P**	0.882±0.016	0.916±0.001	0.911±0.001
	R	0.724±0.021	0.859±0.002	0.892±0.001
	F	0.794±0.019	0.886±0.001	**0.901±0.000**
abandoned_box.oni.	**P**	0.960±0.001	0.887±0.000	0.887±0.000
	R	0.651±0.013	0.837±0.000	0.836±0.000
	F	0.772±0.007	**0.861±0.000**	**0.861±0.000**
illumination_change.oni	**P**	0.645±0.171	0.812±0.000	0.822±0.000
	R	0.591±0.030	0.882±0.000	0.877±0.000
	F	0.574±0.108	**0.846±0.000**	0.848±0.000
interaction.oni	**P**	0.894±0.008	0.719±0.009	0.719±0.009
	R	0.640±0.004	0.643±0.016	0.642±0.016
	F	**0.742±0.003**	0.672±0.010	0.672±0.010
Total	**P**	0.845±0.019	0.834±0.008	0.835±0.007
	R	0.652±0.003	0.805±0.012	0.812±0.013
	F	0.721±0.010	0.816±0.010	**0.821±0.010**

Figure 10. Overall results for foreground segmentation comparing colour and depth information

The following Figure presents sample results for the sequence *illumination_change.oni*

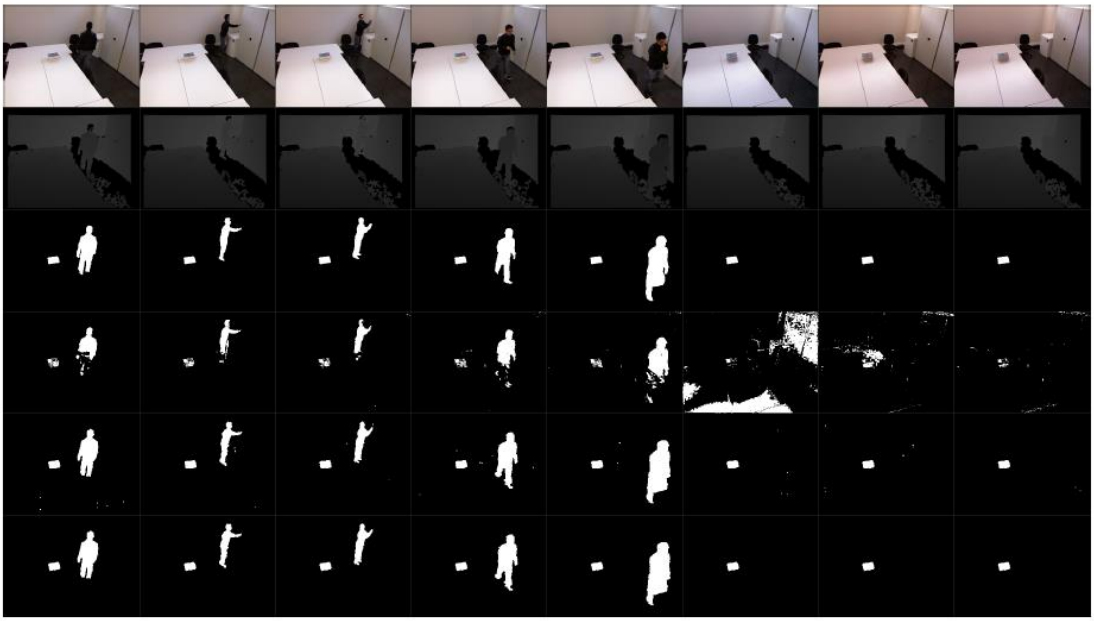


Figure 11. Overall results for foreground segmentation comparing colour and depth information

The following Table presents the overall results for the generated dataset. It can be clearly observed that the improved system efficiently uses depth to increase both the precision and accuracy of the detection of abandoned and stolen objects.

	Sistema inicial						Sistema mejorado					
	TP	FP	FN	P	R	F	TP	FP	FN	P	R	F
Stolen_box-a	1	0	0				1	0	0			
Stolen_box-b	0	0	1				1	0	0			
abandoned_box-a	1	0	0				1	0	0			
abandoned_box-b	0	0	1				1	0	0			
abandoned_box-c	0	0	1				1	0	0			
Stolen_bin	1	0	0				1	0	0			
abandoned_bin	1	0	0				1	0	0			
Meeting_room-carpeta	1	0	0				1	0	0			
Meeting_room-maletín	1	0	0				1	0	0			
Meeting_room-funda	1	0	0				1	0	0			
Meeting_room-móvil	0	0	1				0	0	1			
Illumination_change	0	0	1				1	0	0			
Total	**7**	**0**	**5**	**1**	**0.58**	**0.73**	**11**	**0**	**1**	**1**	**0.92**	**0.96**

Figure 12. Overall results for abandoned and stolen object detection comparing the initial system (only colour) and the improved one (using depth information).

An example is presented in the following Figure where it can be observed that the depth-based system is able to detect the abandoned and stolen objects.

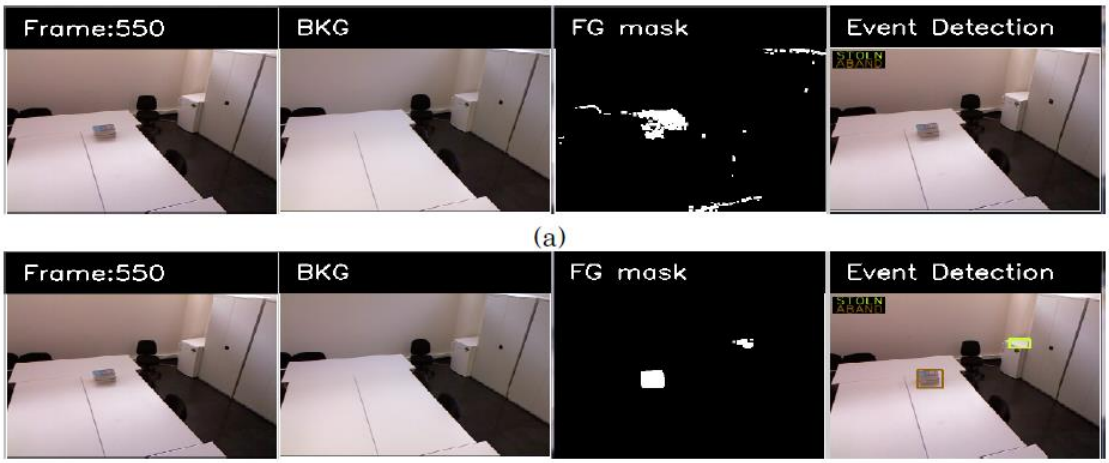


Figure 13. Example for abandoned and stolen object detection using the initial system (top row) and the improved one (bottom row).

2.2. Part-based object recognition through neural-oriented modelling

2.2.1. Approach description

We have developed a neural-oriented strategy for part-based object recognition. Starting only from colour images and depth estimations from the Kinect technology, we focus in the recognition of objects in severe-occlusion and clutter scenarios. To face this scenario, objects are split in successively coarser region-partitions with each region representing a part of the object from which it was extracted. For the characterization of these parts, two new regional descriptors are proposed: R-DAISY and R-SHOT. The former encapsulates luminance and depth information inside a region with a DAISY-like [6] organisation, whereas the latter arranges surrounding normals and colours of three dimensional singular-points in a SHOT-like [7] scheme. Their novelty relies in the use of a size-and-shape-variable description support which is automatically defined by the object part itself. So-obtained descriptions are self-

organised in a single neural structure by an unsupervised learning process. This structure allows to automatically discover relations between the object's parts. The information of object part's is then encoded in a distribute fashion by a set of signs, each corresponding to the response of the neural structure to an object part description. Results show that the approach achieves promising results in the recognition of severe-occluded objects relying only in Kinect-captured data and using a very small set of training instances—2-to-8 short-varied Kinect-captured views per object to recognize. To our knowledge, there are no previous approaches facing this kind of scenario.

2.2.2. Analysis stages

In its stage-flow representation, the proposed approach follows that of a classic recognition system in computer vision. For a previously segregated object, it consists of: a characterisation stage, which is common for both the training and test stages, a training stage, and a test stage.

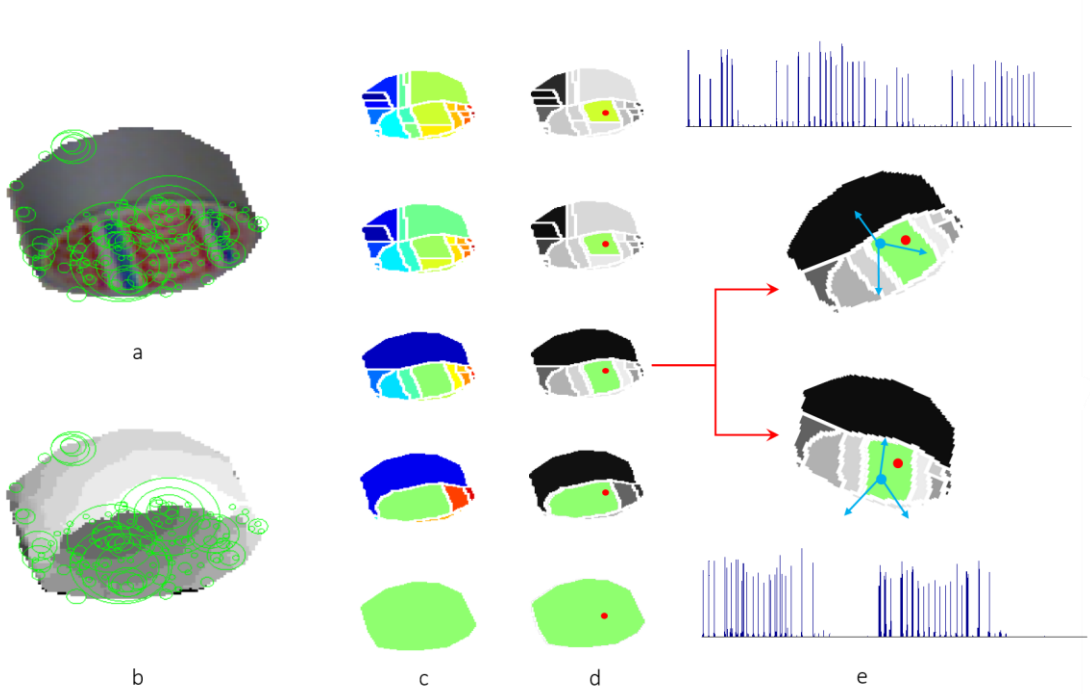


Figure 14. Characterization stage. Singular-points are extracted by means of a scale-space analysis of the chroma and depth information. Points are shown on the colour (a) and depth (b) images of an instance of the *tape* object: the bigger the radio the bigger the scale at which these are detected. The object is then partitioned (c) at several coarseness levels (5 in the figure, from top to bottom). A given pixel (the red dot in column d) may belong to a different region at each coarseness level. In one of the proposed characterization strategies, descriptions of each of these regions are locally-aligned several times, one per each singular-point in the region (blue dots with associated local reference frames in column (e), only two of them are shown to ease visualization), leading to a different object description vector (blue graphs in column (e)).

Characterization stage. In this study, the characterisation process is performed on a regional basis and locally-aligned according to singular-points. Descriptions are extracted around object's singular-points which, as generally accepted, are the most repeatable object's evidences in the presence of moderate object rotations and scale changes. In order to characterise these singular-points, a description support area around them should be used. In existing two and three-dimensional descriptors of singular-points, this area is fixed for every singular-point, both in its shape and in its size. For singular-points extracted near the object boundaries, as the support area may spread beyond the object extent, their description may include information of surrounding objects or even of the scene background. To solve this problem we propose to

restrict the support area to the object's boundaries. In order to perform this automatically, we rely in a region-segmentation approach and use, for the description, information of just the pixels (or voxels) belonging to the region containing the singular-point. The process is depicted in Figure 14.

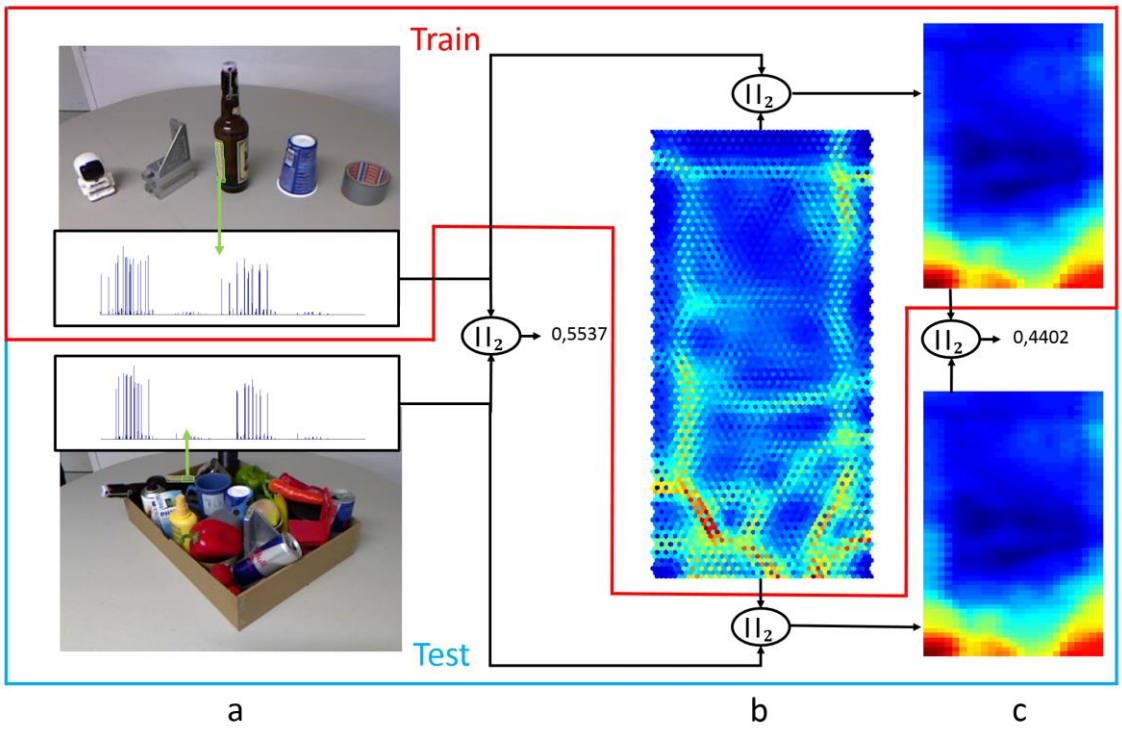


Figure 15. Train / Test stage. After the characterization stage description vectors are generated (a). Training description vectors from different objects, different object's views and different object partitions and local allignetions are used to create the SOM (b). An excitation pattern is obtained by comparing a test description vector against each neuron's weight vector in the SOM (c). Note that whereas the description itself is robust (to some degree) to object rotations (as in this case both object parts share a common reference) and different partitions, the euclidean distance between the responses is even lower than the equivalent computed between the descriptions themselves. Qualitatively, in the example, the excitation patterns are indeed quite similar even when they represent different oriented partitions at different level of coarseness of one of the labels of the bottle object.

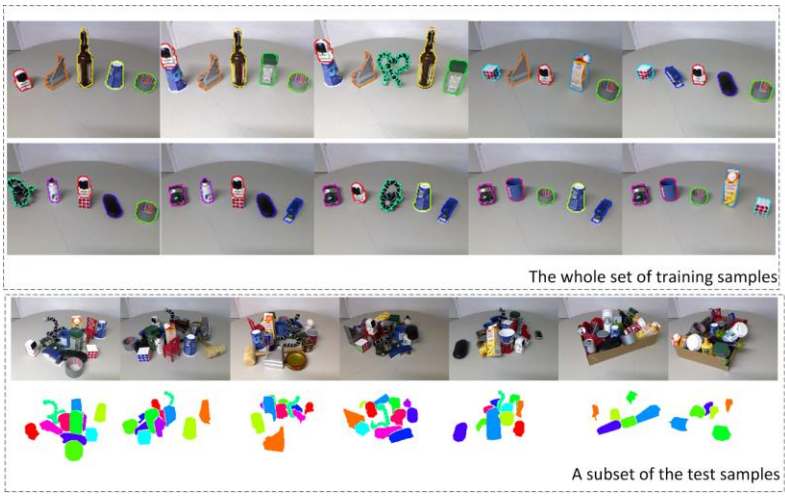
Train stage. All of these so-extracted data for all the training object instances need to be organised and grouped in order to find common descriptions for common object parts. There might be thousands of object descriptions for each object instance and, due to the nature of the characterisation process, different parts of the same object might be described completely different. In order to manage this in an automatic fashion we rely in an unsupervised learning and organising process which we expect: to arrange highly dimensional data in a manageable structure; to store common descriptions together; and to generate classification boundaries in a (highly) multi-dimensional description space. From this training process arises a sheet-like two-dimensional neural map, with each neuron representing—through a description vector (a neuron's weight vector) of the same dimension as those used for training the map—similarly described object *evidences*, which are expected, but not forced, to be from the same object part under the same locally-alignment. Furthermore, similarity relations between each neuron and its neighbouring neurons are also returned and accounted to place neurons associated to the same modality—i.e. representing similar *evidences*—close in the two-dimensional representation space. Observe that the knowledge associated to an object class can be stored at several non-

adjacent neurons in this representation space, each neuron representing a different part (or a different part alignment) of the object. A particularity of the sketched learning process is that, as it is unsupervised, the knowledge arrangement only relies in the description vectors. The process is depicted in the upper part of **Figure 15**,

Test stage. A straightforward way to measure the likelihood of a test description being part of a trained object, would consist of comparing it with all the trained descriptions (description-to-description). However, based on the described neural organisation, an alternative would be to search for the best neuron-to-description association and tag the instance with the neuron's label (description-to-model) in a code-book-like scheme. Finally, a further alternative would be to evaluate the similarity of the two excitation patterns obtained when comparing the model with a trained description and with the test description (pattern-to-pattern). Note that the first option does not rely in any model, the second one is limited to the recognition of the objects observed during the model training stage, whereas only the third allows the introduction of new training descriptions without the requirement of retraining the model, as just the response of the model is required. As our aim is to recognize objects as full entities, and the approach is based on a regional description basis, a region-to-object association is then required. In the proposed approach we follow a labelling procedure: every test description is scored by pattern-to-pattern comparison with the trained descriptions. This process results in a scoring vector which stores the likelihood of such test description being a representation of each of the trained objects. All the test descriptions that involve an object-point are jointly considered to obtain the likelihood of the object-point being part of a trained object. The process is depicted in the lower part of Figure 15,

2.2.3. Experimental evaluation

Dataset description. In order to evaluate the discrimination capability of the selected regional features and the ability of a set of signs to represent an object's knowledge, we propose to evaluate our proposed technique with the challenging dataset presented in [8]. The training part of the dataset is composed of 50 instances of 14 objects distributed in 10 images. In practice, this entails that there are only 2-to-8 short-varied samples per object for the training stage (see **Figure 16**). The testing part includes a total of 448 object instances of the same 14 objects distributed in 75 images, which also include many other unlabelled—and hence untrained—objects. The 14 objects are quite varied in appearance, going from highly textured—*glass*—, through middle textured—*astronaut, carton of juice, cup, toy car, rubik cube*—to untextured or flat objects—*case,stapler*—, these also including metallic objects—*metal piece, spray*—. The dataset also comprises a highly deformable object—*tripod*—, two crystal objects—*bottle, jar*—and a non-compact object—*tape*—. Depth information—due to the Kinect capture technology—is missing in some areas at the crystal: *bottle, jar* and the metallic objects: *metal piece, spray*. In our experiments we have used an interpolation method relying in a 8-connected spring metaphor to estimate missing depths.



Object name	Color code	# instances @ training	# instances @ test
Astronaut		8	26
Metal piece		4	43
Bottle		3	16
Glass		4	52
Tape		7	36
Jar		2	37
Tripod		3	28
Rubik cube		3	22
Carton of Juice		2	51
Stapler		4	18
Case		3	34
Spray		2	34
Toy car		3	15
Cup		2	36

Figure 16. The dataset analysed—presented in [8]—. First row left side: the whole training set of samples. Note that a very few number of samples are available for training. Additionally, in some cases the samples are captured from the same point-of-view. Right side: the objects classes to recognise, colour identification and number of training and test instances per object. Second row: Some frames from the test dataset and associated ground-truth, objects are severely occluded and even placed inside a box.

This dataset is, to our knowledge, first used for the evaluation of an object recognition system—in [8] was used instead for detection of saliency cues—. In our opinion, this is due to a couple of challenging factors that disregards its use: i) as aforementioned, the dataset only contains 10 frames where the objects appear isolated—see first row on the left part of **Figure 16**—and ii) in the test scenario the trained objects are severely occluded by objects unobserved in the training stage. Furthermore, at some of the frames, the objects are placed inside a box—some examples are shown in the second row of **Figure 16**—. However, this dataset perfectly adapts to the target scenario that we consider, as the training set is small, occlusions are varied and natural—not synthetically generated—and object's spatial location is annotated, which allows to bypass the segregation and then evaluate the recognition system independently of this stage.

	Astronaut	M. Piece	Bottle	Glass	Tape	Jar	Tripod	R. Cube	C. of Juice	Stapler	Case	Spray	Toy car	Cup	U
Astronaut	0.70	0.01	0.00	0.00	0.23	0.03	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
M.Piece	0.38	0.28	0.00	0.01	0.14	0.10	0.00	0.02	0.06	0.00	0.00	0.00	0.00	0.00	0.00
Bottle	0.14	0.03	0.57	0.00	0.03	0.05	0.11	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.03
Glass	0.04	0.02	0.00	0.88	0.01	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.00
Tape	0.11	0.05	0.01	0.01	0.76	0.01	0.00	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.02
Jar	0.11	0.02	0.12	0.00	0.03	0.55	0.00	0.00	0.01	0.00	0.05	0.00	0.11	0.00	0.00
Tripod	0.06	0.01	0.00	0.00	0.02	0.01	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
Rubik Cube	0.03	0.03	0.00	0.07	0.15	0.07	0.00	0.59	0.01	0.00	0.01	0.00	0.03	0.00	0.00
Carton of Juice	0.06	0.02	0.15	0.00	0.12	0.17	0.00	0.01	0.46	0.00	0.00	0.00	0.00	0.00	0.00
Stapler	0.01	0.09	0.00	0.02	0.01	0.06	0.00	0.00	0.00	0.79	0.00	0.00	0.01	0.00	0.01
Case	0.02	0.01	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.87	0.00	0.05	0.00	0.00
Spray	0.43	0.03	0.05	0.00	0.10	0.05	0.00	0.01	0.03	0.00	0.00	0.29	0.00	0.00	0.00
Toy car	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.26	0.00	0.73	0.00	0.00
Cup	0.00	0.01	0.00	0.15	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.74	0.02

Figure 17. Quantitative results. Confusion matrix for the proposed approach.

Results. Quantitative results, in the shape of the system's confusion matrix for the whole approach are rendered in **Figure 17**, whereas some qualitative operation examples are also included in **Figure 18** for completeness. Results are computed on a per object-point basis, i.e., the diagonal of the confusion matrix stands for the fraction of object-points of a given object class correctly classified as being part of that class.

Figure 18. Qualitative results. Test images (first row) and associated ground-truth (second row). Recognition achieved by the proposed system (third row).

3. Conclusions and Future Work

In the first contribution (depth-based abandoned object detection), it has been showed that depth information significantly improves the performance of foreground segmentation since it does not present false positives. However, the opposite situation, where false negatives appear, is more frequent since it depends on the physical properties of the objects in the scene. Combining colour and depth information seems to face such limitation and improve the results in most of the experiments performed. Moreover, the obtained foreground masks do not need to undergo additional processing for shadow removal since depth data is inherently shadow-free. As limitations, the proposed approach requires to detect small blobs using colour information in order to include them in the initial depth mask. Therefore, determining when a blob is small is not straightforward and depends on the application. As future work, the following areas can be explored: inclusion of negligible-depth objects in the foreground mask, efficient updates of the depth-based background model for long term analysis and the operation with dynamic cameras with variable field of view such as Pan-Tilt-Zoom cameras.

In connection with the part-based object recognition system, a novel approach that, based on psychophysical considerations, assumes that independent to appearance and occlusions object recognition relies in part-based descriptions based on masked local features of such parts has been presented. The system organises evidences from a very small collection of objects—without the use of CAD models—in a single and relatively simple neural structure, leaving space for further knowledge. The approach has proven to provide promising results in the recognition of severely occluded objects by using a very small and short-varied subset of

objects in the training stage. In particular, reported results show the benefits of including a regional-version of a recent state-of-the-art 3D descriptor in the proposed neural-framework. Furthermore, the benefits of using distributed encoding in the faced scenarios have been also experimentally evaluated with success. Our future work in this vein is devoted to: i) enhance the system recognition capability by grouping signs in order to pave the road to its extension to object's class recognition, i.e. different entities (two different cups) of an object class (*cup*); ii) improve the designed local descriptor by exploring new schemes for geometrical organisation or by its substitution for alternative descriptors; and iii) further explore the benefits of distribute encoding by establishing, through evaluation, the representation limits of the constructed neural map.

References

[1] O. Steiger A. Cavallaro and T. Ebrahimi. Semantic video analysis for adaptive content delivery and automatic description. Circuits and Systems for Video Technology, IEEE Transactions on, pages 15(10):1200–1209, Oct 2005.

[2] Massimo Piccardi Rita Cucchiara, Costantino Grana and Andrea Prati. Detecting moving objects, ghosts, and shadows in video streams. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 25:1337–1342, 2003.

[3] P. Salembier and J. Ruiz. On filters by reconstruction for size and motion simplifcation. In Proc. of Int. Symposium in Mathematical Morphology, pages 425–434, 2002.

[4] José M. Martínez Álvaro Bayona, Juan C. SanMiguel. Stationary foreground detection using background subtraction and temporal difference in video surveillance. In Image Processing (ICIP), 2010 17th IEEE International Conference on, pages 4657–4660, Hong Kong (China), 2010.

[5] L. Caro J.C. SanMiguel and J.M. Martínez. Pixel-based colour contrast for abandoned and stolen object discrimination in video surveillance. In Electronics Letters, pages 86–87, 2012.

[6] Tola, E., Lepetit, V., Fua, P., May 2010. DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (5), 815–830.

[7] Salti, S., Tombari, F., Stefano, L. D., 2014. Shot: Unique signatures of histograms for surface and texture description. Computer Vision and Image Understanding 125, 251–264.

[8] Potapova, E., Zillich, M., Vincze, M., 2011. Learning what matters: Combining probabilistic models of 2d and 3d saliency cues. In: Crowley, J., Draper, B., Thonnat, M. (Eds.), Computer Vision Systems. Vol. 6962 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 132–142.