VOL. 12, NO. 11, JUNE 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

GAUSSIAN PROCESS APPROACH FOR VISUALLY IMPAIRED PEOPLE TO IDENTIFY OBSTACLES BASED ON EUCLIDEAN DISTANCE MEASURE

M. Karthikeyan[1], Joseph Henry[2] and K. Rajan[3]
[1]Research Scholar, Vel Tech Dr. RR Dr. SR Technical University, Chennai, India
[2]Professor, School of Electrical Engineering, Veltech Dr. RR Dr. SR Technical University, Chennai, India
[3]Professor and Dean, Department of Electrical & Electronics Engineering, Veltech, Chennai, India
E-Mail: karthickm37@gmail.com

ABSTRACT

This paper introduces a new portable camera-based method for helping blind people to recognize indoor objects. Unlike state-of-the-art techniques, which typically perform the recognition task by limiting it to a single predefined class of objects, we propose here a completely different alternative scheme, defined as coarse description. It aims at expanding the recognition task to multiple objects and, at the same time, keeping the processing time under control by sacrificing some information details. The benefit is to increment the awareness and the perception of a blind person to his direct contextual environment. The coarse description issue is addressed via two image multilabeling strategies which differ in the way image similarity is computed. The first one makes use of the Euclidean distance measure, while the second one relies on a semantic similarity measure modeled by means of Gaussian process estimation. To achieve fast computation capability, both strategies rely on a compact image representation based on compressive sensing. The proposed methodology was assessed on two indoor datasets representing different indoor environments. Encouraging results were achieved in terms of both accuracy and processing time.

Keywords: assistive technologies, blind people, compressive sensing (CS), Gaussian processes (GPs), indoor scene description.

1. INTRODUCTION

A recent revision of visual impairment definitions in the international statistical classification of diseases, carried out in 2011, has revealed that visual acuity and performance are categorized according to one of the following four levels, namely, normal vision, moderate, severe, and blindness [3]. Regardless of such terminologies, visual impairment in general and blindness in particular, as any other pathology, have their adverse impacts as regards to both physical and moral aspects. In spite of the remarkable medical efforts being dedicated to cope with vision disability, the big prospective leap to full sight recovery has not yet been met. Nonetheless, supportive solutions could be a means toward a partial recovery. Consequently, assistive rehabilitation technologies have been finding their way for satisfying such need to a reasonable extent.

In quest of visual disability rehabilitation, several prototypes and designs have been proposed so far, and have dealt with different issues. From the literature, one can deduce that the overwhelming majority of the contributions could be highlighted under one of the two categories confined to navigation (i.e., by allowing more freedom in terms of mobility, orientation, and obstacle avoidance) and recognition (i.e., by providing the blind person with information related to the nature of objects encountered in his/her context). Regarding the navigation issue, which has been devoted the biggest part of interest as compared with the recognition aspect, different contributions have been carried out, and generally two main groups are considered in the literature.

The first one relies on active devices, in which case some sort of signal or beam is sent and subsequently received back and the duration consumed between both processes defined as time of flight is exploited, as proposed for instance in [4]-[8]. However, the main drawbacks of such devices are their size on the one hand and their power consumption on the other hand, which reduce their suitability for daily use by a visually impaired individual. However, thanks to the ever-increasing interest witnessed in computer vision, such issues have become retractable. As a consequence, recent works are oriented toward computer vision, such as for instance [9]–[12]. As for the recognition aspect, relatively few contributions could be found in the literature and are mostly computer vision based. In [13], for instance, a banknote recognition system for the blind was proposed. It relies basically on the well-known speeded-up robust features (SURF). López-de-Ipiña*et al.* [14] suggested a supported supermarket shopping, which incorporates navigational tips for the blind person through Radio frequency identification technology, and camera-based product recognition via Quick response codes placed on the shelves. A product barcode detection as well as reading was developed in [15]. In [16], a travelling assistant was proposed. It takes advantage of the text zones depicted in the frontal side of buses (at bus stops) for further extraction of information related to line number and the coming bus. The system processes a given image acquired by a portable-camera, and then notifies the outcome to the user vocally. In another computer vision-based contribution [17], assisted indoor staircases detection (within 1-5 m ahead) was suggested. Also proposed in [18] is an algorithm intended to help visually impaired people to detect as well as read text encountered in natural scenes. Yang and Tian [19] proposed to assist blind

persons to detect doors in unfamiliar environments. Assisted indoor scene understanding through indoor signage detection and recognition was also considered in [17], through the use of the popular scale invariant feature transform.

Accordingly, from the state of the art reported so far, it is possible to make out that object detection and/or recognition for the blind is approached in a class-specific manner. In other words, all the contributions tend to emphasize on the recognition of one specific category of objects. Such strategy (i.e., focusing the interest on one class of objects), despite its effectiveness, conveys useful but limited information for the blind person. By contrast, extending the interest to recognizing multiple different objects at once can be looked at as an alternative approach to make the recognition task more generalized and informative. It is also aiming at bringing closer the indoor scene description to the blind person, yet fostering his/her imagination. This is, however, not an easily achievable task due to the number of algorithms that would be invoked simultaneously (in case of setting up one algorithm per specific object), and may result in an unwanted high processing overcharge, thus making a real-time or even a quasi-real-time implementation infeasible.

In the general computer vision literature, several works dealing with multiobject recognition can be found in [20]-[24]. In [20], for instance, a novel approach for semantic image segmentation is investigated. The proposed scheme relies on a learned model, which derives benefits from newly proposed features, termed texture-layout filters, incorporating texture, layout, and context information. Presented in [21] is a scalable multiclass detector, in which a shared discriminative codebook of feature appearances is jointly trained for all object classes. Subsequently, taxonomy of object classes is built based on the learned sharing distributions of features among classes, which is thereupon taken as a means to lessen the cost of multiclass object detection. Following a scheme that combines local representations with region segmentation and template matching, in [22], an algorithm for classifying images containing multiple objects is presented. Generative model-based object recognition is proposed in [23]. It makes use of a codebook derived from edge based features. Pantofaru*et al*. [24] introduces an object recognition approach that starts from a bottom-up image segmentation and analyzes the multiple segmentation levels of the image. In general, it emerges that most of the contributions deal with the multirecognition issue as an image segmentation problem and propose solutions not particularly adapted to the context of blind assistance because of tight time processing requirements.

In this regard, this paper proposes an alternative approach meant to solve the problem of multiobject detection in images acquired in indoor environments. The underlying idea is to trade computation time with object information details, such as the position of the objects within the field of view and their number (i.e., number of times a same object appears in the image). In other words, we propose a new way to perceive the objects in the surrounding environment by means of a coarse but broad and fast description of the scene. The proposed approach will be implemented through image multilabeling. Given a query image (acquired by a portable chest-mounted camera), as a first step it is represented as a compact sequence of coefficients by means of the compressive sensing (CS) representation [23], [24]. To fulfill the multilabeling task, the most resembling images are picked up from a library of images (constructed offline), and then combined to identify the objects characterizing the query image. As to cope with image similarity assessment, in alternative to the Euclidean distance, we present a semantic-based similarity measure accomplished through a statistical prediction model as described further. The algorithms were tested on datasets corresponding to two different public sites and revealed encouraging results in terms of accuracy and processing time.

The remainder of this paper is outlined as follows. Section II presents the basic insight underlying the image multilabeling procedure. Section III describes the CS image representation technique. Section IV presents the proposed semantic similarity measure. Experimental part is conducted in Section V. Ultimately; we summarize the conclusion in Section VI.

2. COARSE IMAGE DESCRIPTION

The purpose in this project is to coarsely describe a given camera-grabbed image of an indoor scene, whose description consists of checking the presence/absence of different objects of interest (determined a priori) and turns out to convey the list of the objects that are most likely present in the scene regardless of their position within the image. The reason behind such a framework is to enrich the perception and the imagination of the blind person regarding the surrounding environment. The proposed image multilabeling process is shown in Figure-1.

The underlying insight as hinted earlier is to compare the considered query image with an entire set of training images. These lasts are captured and stored offline along with their associated binary descriptors, which encode their content, as shown in Figure-2.

The binary descriptors of the k most similar images are considered for successive fusion to multilabel the given query image. This fusion step, which aims at achieving better robustness in the decision process, is based on the simple majority-based vote applied on the k most similar images (i.e., an object is detected in the query image only if, amongst the k training images, it exists once for $k = 1$, at least twice for $k = 3$, and at least thrice for $k = 5$). For that purpose, each training image in the library earns its own binary multilabeling vector (or simply image descriptor), which feeds the fusion operator.

The routine for establishing such vector for a given training image is to visually check the existence of each object within a predefined list in the image. If an object exists within a given depth range ahead assessed by visual inspection of the considered training image (e.g., 4 m), then a 1 is assigned to its associated bin in the vector, otherwise a 0 value is retained, as shown in Figure-3.

3. SPARSE IMAGE REPRESENTATION

As aforesaid, the underlying idea is to multilabel a given query image by fusing the content of the most similar training images in the library. Hence, the way the matching is performed represents a decisive part. This implies the adoption of two main ingredients: 1) a suitable image

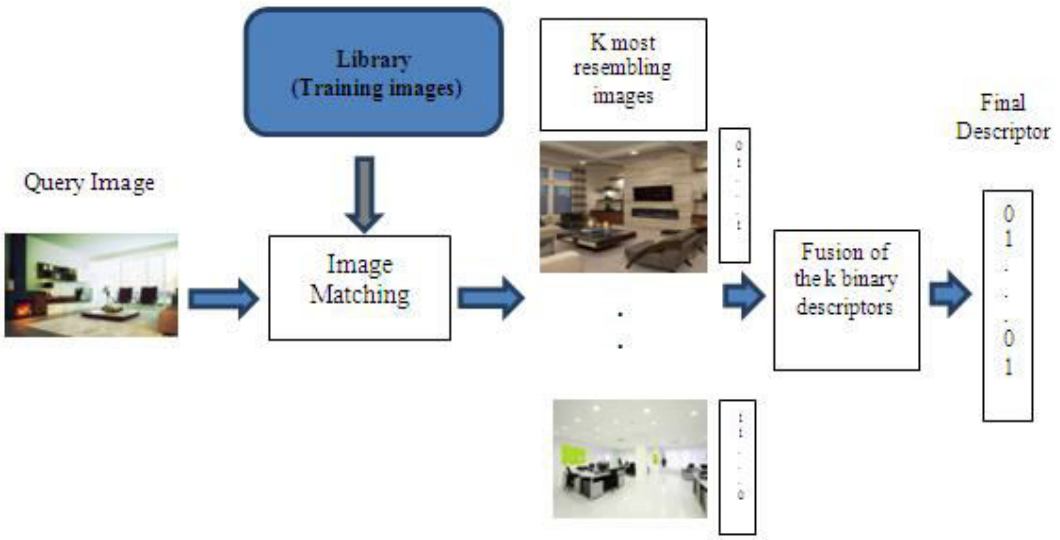

Figure-1. General block diagram of the proposed framework for coarse scene description (image multilabeling).

Binary descriptor	Objects	Present?
1	Stair Door	Yes
0	Pillar	No
0	ATM machine	No
0	Chairs	No
1	External Doors	Yes
0	Bins	No

Figure-2. Relationship between the binary descriptor and the predefined set of objects.

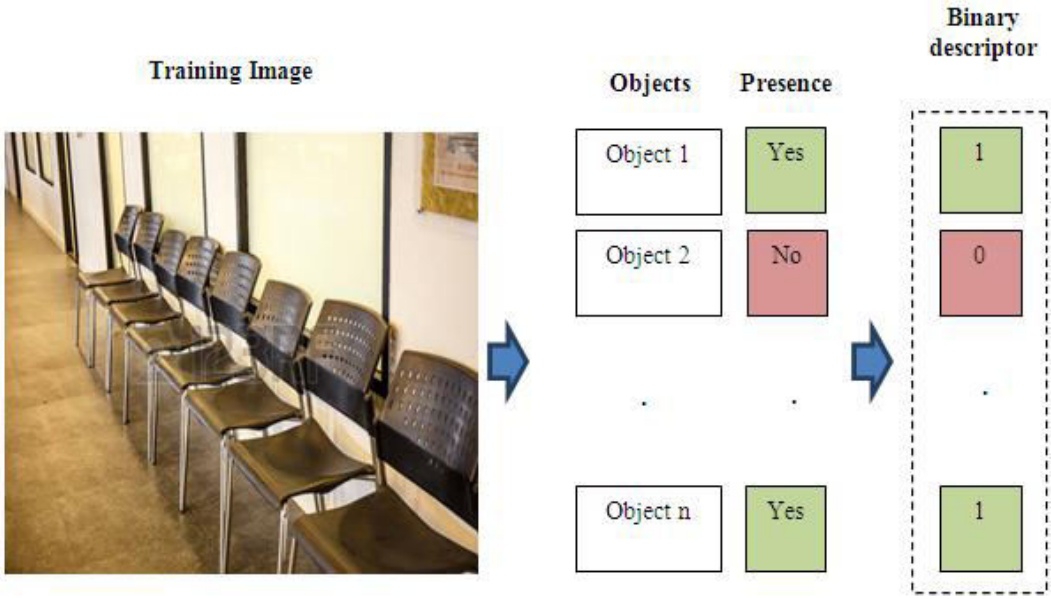

Figure-3. Procedure for constructing binary image descriptors for the training images.

representation and 2) a similarity measure. Regarding the former ingredient, there is a need for an appropriate tool to represent the images dealt with in a compact way for being able to achieve fast image analysis. Among recent possible compact representations is the CS theory [1], [2], which has gained an outstanding position and become a significant tool in the signal processing community. In the following, we will, respectively, provide foundational details outlining the main CS concepts, and describe how it is exploited in our work for compact image representation.

A. Compressive Sensing Theory

CS, also known as compressive sampling, compressed sensing or sparse sampling, was recently introduced in [1] and [2]. CS theory aims at recovering an unknown sparse signal from a small set of linear projections. By exploiting this new and important result, it is possible to obtain equivalent or better representations using less information compared with traditional methods (i.e., lower sampling rate or smaller data size). CS has been proved to be a powerful tool for several applications, such as acquisition, representation, regularization in inverse problem, feature extraction, and compression of high-dimensional signals, and applied in different research fields, such as signal processing, object recognition, data mining, and bioinformatics [25]. In these fields, CS has been adopted to cope with several tasks like recognition [26]-[28], image super-resolution [29], segmentation [30], denoising [31], inpainting and reconstruction [32], [33], and classification [34]. Note that images are a special case of signals which hold a natural sparse representation, with respect to fixed bases, also called dictionary (i.e., Fourier and wavelet) [35].

CS is, thus, a way to obtain a sparse representation of a signal. It relies on the idea to exploit redundancy (if any) in the signals [1], [2]. Usually, signals like images are sparse, as they contain, in some representation domain, many coefficients close to or equal to zero. The fundamental of the CS theory is the ability to recover with relatively few measurements $V = D \cdot \alpha$ by solving the following $L0$-minimization problem:

$$min \; || \alpha ||_0 s.t. \; V = D \, . \, \alpha \qquad (1)$$

where D is a dictionary with a certain number of atoms (which in our case, are images converted into vectors), V is the input image (converted into vector) which can be represented as a sparse linear combination of these atoms, α is the set of coefficients intended as a compact CS-based representation for the input image V. The minimization of $\| \, . \, \|_0$, the $L0$-norm, corresponds to the maximization of the number of zeros in α, following this formulation: $\| \alpha \|_0 = \# \{ i = \alpha_i \neq 0 \}$. Equation (1) represents an NP-hard problem that means that it is computationally infeasible to solve. Following the discussion in [36], it is possible to simplify the evaluation of (1) in a relatively easy linear programming solution. They demonstrate that, under some reasonable assumptions, minimizing L_1-norm is equivalent to minimizing $L0$-norm, which is defined as $\| \alpha \|_1 \sum_i |a_i|$. Accordingly, it is possible to rewrite (1) as

$$min \; || \alpha ||_1 s.t. \; V = D \, . \, \alpha \qquad (2)$$

In the literature, there exist several algorithms for solving optimization problems similar to the one expressed in (2). In the following, we briefly introduce an effective algorithm called stagewise orthogonal matching pursuit (StOMP) [37], which will be used in our work. By contrast to the basic orthogonal matching pursuit (OMP) algorithm, StOMP involves many coefficients at each stage (iteration), while in OMP only one coefficient can be involved. In addition, StOMP runs over a fixed number of stages, whereas OMP may take numerous iterations. Hence, StOMP was preferred in our work because of its fast computation capability.

B. CS-based image representation

The use of the CS theory for image representation in our work is thus motivated by its capability to concisely represent a given image. For such purpose, a bunch of N_C learning images representing the indoor environment of interest is first acquired. All images (if in Red, Green, and Blue format) are converted in grayscale and into vectors. Their column-wise concatenation forms the dictionary D (composed of N_C atoms). Given a query image V, its compact representation α (whose dimension is reduced to the number of learning images) is achieved by means of the procedure summarized below:

Step 1: Consider an initial solution $\alpha_0 = 0$, an initial residual $r_0 = V$, a stage counter s set to 1, and an index sequence denoted as $T_1,\ldots,T_S$, which contains the locations of the nonzeros in α_0.

Step 2: Compute the inner product between the current residual and the considered dictionary D

$$C_S = D^T .r_{(S-1)} \quad (3)$$

Step 3: Perform a hard thresholding to find out the significant nonzeros in $C_S$by searching for the locations corresponding to the large coordinates J_S

$$J_S = \{j: C(j) > t_S\sigma_S \quad (4)$$

where σ_S represents a formal noise level and t_S is a threshold parameter taking values in the range $2 \leq t_S \leq 3$.

Step 4: Merge the selected coordinates J_S with the previous support

$$T_S = T_{S-1} U J_S \quad (5)$$

Step 5: Project the vector V on the columns of D that correspond to the previously updated T_S. This yields a new approximation α_S

$$(\alpha S)TS = (D^t_{T_S})\text{-}1\ D^t_{T_S} V \quad (6)$$

Step 6: Update the residual according to $r_S = V - D \cdot \alpha s$.

Step 7: Check whether a stopping condition (e.g., smax= 10) is met. If so, $\alpha_S$is considered as the final solution. Otherwise, the stage counter s is incremented and the next-stage process is repeated starting from Step 2.

The procedure for generating the vector of CS coefficients is shown in Figure-4.

4. SEMANTIC IMAGE SIMILARITY MEASURE

As mentioned above, the second ingredient to be adopted for image matching is the similarity measure. In this project, we will interpret the term similarity in two different ways. The first one is the distance between two images in a given image domain representation, in our case in the CS coefficient domain. For measuring the distance, we will make use of the well-known Euclidean distance. The second way of interpretation consists to compare the images in a semantic domain. This means that two images are semantically close if they contain the same objects, regardless of the apparent image resemblance. To that end, we propose in this project a semantic-based framework for quantifying the similarity between images. Its underlying

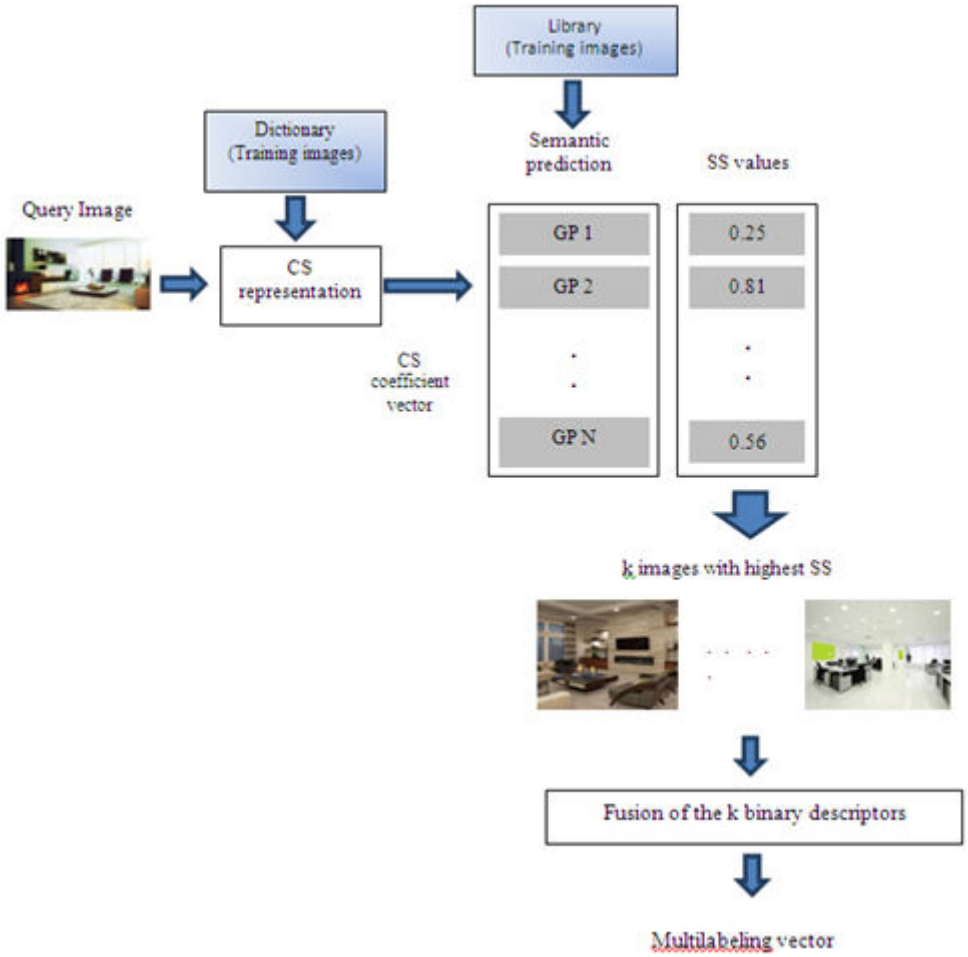

Figure-4. Flowchart of the proposed SSCS image multilabeling strategy.

idea is to go through a semantic similarity predictor, learned *a priori* on a set of training images to predict the extent up to which two given images are semantically close. Among the variety of existing predictors, we will opt for the Gaussian process (GP) regression model because of its good generalization capability and short processing time. In the following sections, more details about the proposed semantic similarity prediction and the GP regression are provided, respectively.

A. Semantic similarity

Given two images I_1 and I_2 together with their corresponding binary descriptors b_1 and b_2, we define the quantity $SS_{I1,I2}$ as the semantic similarity between I_1 and I_2. In particular, this measure expresses the ratio inclusion of I_2 in I_1, that is the number of objects of I_2 (represented as ones in b_2) present also in I_2 (i.e., still represented as ones in b_1). Hence, the larger the $SS_{I1,I2}$ the (semantically) closer I_2 to I_1. Mathematically, it is expressed by

$$SS_{I_1,I_2}=\frac{\sum_{i=1}^{N} b_1(i).\ b_2(i)}{\sum_{i=1}^{N} b_1\ (i)} \quad (7)$$

The multilabeling process based on the semantic similarity prediction is articulated over two phases as follows:

1) Training phase: First, compute the SS values between all couples of training images. Then, train as many GP regressors as the number of training images (i.e., N). Each GP regressor will be learned to predict $SS_{Ip,Ii}$, that is the semantic similarity between a given generic image I and the training image $I_p$to which the GP regressor is associated. The supervised training of the pth predictor is performed by giving: 1) in input the CS coefficients corresponding to each training image I_i and 2) in output as target the $SS_{Ip,Ii}$ values (between reference image I_p and each training image I_i).

2) Operational phase: Feed each GP predictor with the CS coefficient vector of the query image I to estimate all $SS_{Ip,Ii}$ values, i.e., the similarity between I and each training images I_p.

Subsequently, pick up the k binary descriptors associated with the training images corresponding to the k highest SS values for successive fusion, and infer the multilabeling of the query image as explained earlier. Figure-5 shows the semantic similarity CS (SSCS) strategy.

B. Gaussian process regression

According to the GP formulation [38]–[40], the learning of a machine is expressed in terms of a Bayesian estimation problem, where the parameters of the machine

are assumed to be random variables which are *a priori* jointly drawn from a Gaussian distribution. In greater detail, let us consider $\mathbf{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ a matrix of input data representing our N training Images, where $\mathbf{x}_i \in \mathbb{R}^{Nc}$ represents a vector of N_c processed features, namely, the N_cCS coefficients associated with the ith training image.

Let also denote $\mathbf{y} = \{\boldsymbol{y}_i\}_{i=1}^{N}$ as the corresponding output target vector, which collects the desired semantic similarity values (between the considered reference image and all the training images). The aim of GP regression is to infer from the set of training samples $\{\mathbf{X}, \mathbf{y}\}$ the function $\psi(\cdot)$ so that $\mathbf{y} = \psi(\mathbf{x})$. This can be done by formulating the Bayesian estimation problem directly in the function space view. The observed values $\mathbf{y}$ of the function to model are considered as the sum of a latent function $\mathbf{f}$ and a noise component ε, where

$$f \sim GP\{0, K(X,X)\} \tag{8}$$

and

$$\varepsilon \sim N(0, \sigma_n^2 \boldsymbol{I}) \tag{9}$$

Equation (8) means that a GP$\{\cdot, \cdot\}$ is assumed over the latent function f, i.e., this last is a collection of random variables, any finite number of which follow a joint Gaussian distribution [39]. $\mathbf{K}(\mathbf{X},\mathbf{X})$ is the covariance matrix, which is built by means of a covariance (kernel) function computed on all the training sample pairs. Equation (9) states that a Gaussian distribution with zero mean and variance $\sigma_n^2$is supposed for the entries of the noise vector ε with each entry drawn independently from the others ($\mathbf{I}$ represents the identity matrix). Because of the statistical independence between the latent function $\mathbf{f}$ and the noise component ε, the noisy observations $\mathbf{y}$ are also modeled with a GP

$$y \sim GP(0, \boldsymbol{K(X,X)} + \sigma_n^2 \boldsymbol{I}) \tag{10}$$

or equivalently

$$p(y|X) = N(0, \boldsymbol{K(X,X)} + \sigma_n^2 \boldsymbol{I}) \tag{11}$$

In the inference process, the best estimation of the output value $f*$associated with an unknown sample $x*$is given by

$$\widehat{f}_* \mid X, y, x_* \sim E\{f_*|X, y, x_*\} = \int f_* p(f_*|X, y, x_*) df \tag{12}$$

From (12), it is clear that, for finding the output value estimate, the knowledge of the predictive distribution $p(f_*|X, y, x_*)$is required. For this purpose, the joint distribution of the known observations $\mathbf{y}$ and the desired function value $f*$ should be first derived. Thanks to the assumption of a GP over $\mathbf{y}$ and to the marginalization property of GPs, this joint distribution is Gaussian. The desired predictive distribution can be derived simply by conditioning the joint one to the noisy observations $\mathbf{y}$ and takes the following expression:

$$p(f_*|X, y, x_*) = N(\mu_*, \sigma_*^2) \tag{13}$$

where

$$\mu_* = k_*^T . [\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \sigma_n^2 \boldsymbol{I}]^{-1} . y \tag{14}$$

$$\sigma_n^2 = k(x_*, x_*) - k_*^T . [\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \sigma_N^2 \boldsymbol{I}]^{-1} . k_* \tag{15}$$

These are the key equations in the GP regression approach. Two important pieces of information can be retrieved from them: 1) the mean μ_*, which represents the best output value estimate for the considered sample according to (12) and depends on the covariance matrix K(X,X), the kernel distances between training and test samples $k*$the noise variance σ_n^2, and the training observations y, and 2) the variance σ_n^2,which expresses a confidence measure associated by the model to the output. A central role in the GP regression model is played by the covariance function $k(x_i , x_j)$ as it embeds the geometrical structure of the training samples. Through it, it is possible to define the prior knowledge about the output function $F(\cdot)$. In this paper, we shall consider the following Matérn covariance function [39]:

$$k(x_i, x_j) = \theta_0 \left(1 + \frac{\sqrt{3}|x_i - x_j|}{l}\right) \exp\left(-\frac{\sqrt{3}|x_i - x_j|}{l}\right) \tag{16}$$

For this covariance function, the hyperparameter vector is given by $\Theta = [l, \theta 0]$. Such vector can be determined empirically by cross validation or using an independent set of labeled samples called validation samples. As an alternative, as it will be done in this paper, the intrinsic nature of GPs allows a Bayesian treatment for the estimation of Θ. For such purpose, one may resort to the type II maximum likelihood estimation procedure. It consists in the maximization of the marginal likelihood with respect to Θ, that is the integral of the likelihood times the prior

$$p(y|X) = p(y|X, \theta) = \int p(y|f, X, \theta) p(f|X, \theta) df \tag{17}$$

with the marginalization over the latent function f. Under a GP regression modeling, both the prior and the likelihood follow Gaussian distributions. After some manipulations, it is possible to show that the log marginal likelihood can be written as [39]

$$\log p(y|X, \Theta) = -\frac{1}{2} y^T . (K(X,X) + \sigma_n^2 I)^{-1} . y - \frac{1}{2} \log|K(X,X) + \sigma_n^2 I| - \frac{n}{2} \log(2\pi) \tag{18}$$

As it can be observed, (18) is the sum of three terms. The first is the only one that involves the target observations. It represents the capability of the model to fit the data. The second one is the model complexity penalty, while the third term is normalization constant. From an implementation view point, this maximization problem

can easily be solved by a gradient-based search routine [39].

5. EXPERIMENT RESULTS

A. Dataset description

The images processed in this paper were acquired by means of a chest-mounted CMOS camera from the IDS Imaging Development Systems; model UI-1240LE-C-HQ with KOWA LM4NCL lens, carried by a wearable lightweight shield, as shown in Figure-6. This last shows the multisensor prototype on which we are working for both guiding blind people and helping them in recognizing objects in indoor environments. The method for the recognition part, which is described in this project, is exploited on demand that is the user has access to the

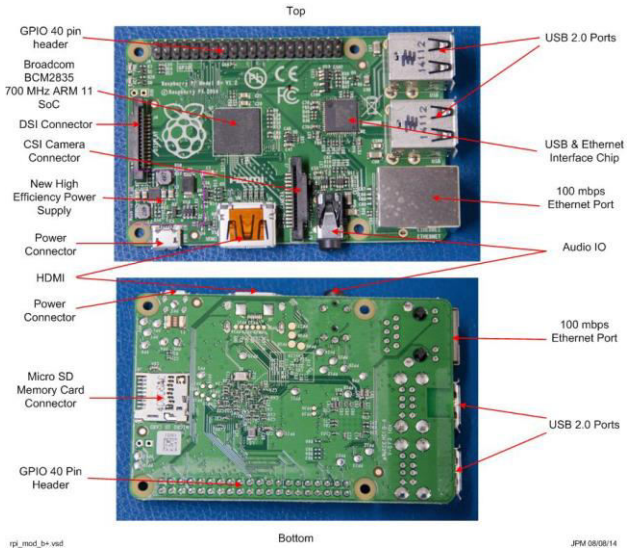

Figure-5. Wearable prototype used for image acquisition.

recognition capability only when he desires it through a vocal instruction. The names of the objects identified within the image extracted from the video stream at the moment of the vocal instruction are communicated by speech synthesis. Work is in progress to integrate all the developed algorithms (including those presented here) in the prototype. Coming back to this project, the image size is 640 × 480 pixels. It is also noteworthy that the images acquired by the portable camera were not compensated for lens distortions, as our method handles the images as a whole and does not extract any feature from within the images.

The collection of images adopted for evaluating the efficiency of the proposed image description method refers to two different buildings in the University. The first set accounts for a total of 181 images acquired in two separate daytimes (morning and evening), which was split into 51 dictionary (learning) images (i.e., exploited to compose the CS dictionary), 58 training images (i.e., for training the GP model), and 72 for testing purposes. The second set is composed of 185 images, divided into 54 CS dictionary (learning) images, 61 training images, and 70 testing images. It is noteworthy that the training images for both datasets were selected in such a way to cover all the predefined objects in the considered indoor environment.

As noted above, a list of objects of interest must be predefined. Thereupon, we have selected the objects deemed to be the most important ones in the considered indoor environments. Regarding the first dataset, 15 objects were considered as follows: External Window, Board, Table, External Door, Stair Door, Access Control Reader, Office, Pillar, Display Screen, People, ATM, Chairs, Bins, Internal Door, and Elevator. Whereas, for the second set, the list was the following: Stairs, Heater, Corridor, Board, Laboratories, Bins, Office, People, Pillar, Elevator, Reception, Chairs, Self-Service, External Door, and Display Screen.

Table-1. Results of proposed strategies obtained on Dataset 1, By varying image resolution and *k* (Number of multilabeling images) value.

		SSCS (Semantic Similarity Compressed Sensing)				EDCS (Euclidean Distance Compressed Sensing)			
Ratio		**1/10**	**1/5**	**1/2**	**1**	**1/10**	**1/5**	**1/2**	**1**
k=1	SEN	80.89	81.64	79.77	79.77	71.53	70.41	69.66	69.66
	SPE	68.14	67.40	66.91	66.54	79.33	79.82	79.82	80.19
k=3	SEN	78.65	78.65	80.52	80.14	65.91	66.66	67.41	68.53
	SPE	69.86	69.61	69.74	69.37	81.54	80.93	81.42	81.91
k=5	SEN	76.02	76.77	76.02	75.65	67.41	67.79	67.79	68.16
	SPE	71.09	70.60	70.47	70.72	82.41	81.91	81.79	82.04

Table-2. Results of proposed strategies obtained on Dataset 2, By varying image resolution and *k* (Number of multilabeling images) value.

		SSCS (Semantic Similarity Compressed Sensing)				EDCS (Euclidean Distance Compressed Sensing)			
Ratio		**1/10**	**1/5**	**1/2**	**1**	**1/10**	**1/5**	**1/2**	**1**
k=1	SEN	75	74.09	75	75	69.18	69.09	70	70
	SPE	73.97	73.73	74.09	74.09	89.03	89.51	90.12	90.12
k=3	SEN	69.54	70.90	70.90	70.45	63.18	62.27	61.36	60.90
	SPE	81.80	82.53	82.65	82.65	87.22	86.98	86.98	87.10
k=5	SEN	68.63	69.09	69.09	68.63	53.18	55	55.90	55.90
	SPE	81.08	81.68	81.92	82.04	89.63	89.39	89.87	89.75

B. Discussion

The efficiency of the proposed framework is expressed in terms of the well-known sensitivity (SEN) and specificity (SPE) accuracy measures defined in [19] and [20]. They express the probability of correct detection of the presence and absence of an object, respectively

$$Sensitivity = \frac{TruePositives}{TruePositives+FalseNegatives} \quad (19)$$

$$Sensitivity = \frac{TrueNegatives}{TrueNegatives+FalsePositives} \quad (20)$$

A worth mentioning fact is that the resolution of the images has a direct influence on the processing time (in particular, in the CS representation phase). Therefore, we have analyzed its impact by running the experiments on four different resolution ratios. The first one is set to the unity (thus, keeping the original 640 × 480 resolution), the second ratio was set to the half (320 × 240 image size), the third one equals to one fifth (128 × 96), and the last one was fixed to one tenth (64×48). The results corresponding to the combination of the k values and the image resolutions regarding both strategies are summarized in Tables 1 and 2, for dataset 1 and dataset 2, respectively.

Considering the results obtained from the first dataset, it comes out that, in overall terms, both the semantic SSCS and Euclidean distance-based compressed sensing (EDCS) methods perform equivalently on an average over the SEN and SPE accuracies. However, the SSCS strategy yields a better SEN, while the EDCS shows a better SPE. As for the second dataset, the EDCS performs slightly better by taking the averages for k = 1. For the other k values, the SSCS outperforms. This is explained by the fact that the EDCS relies on measuring the similarity of the CS coefficients, yet measuring the apparent similarity between the images, which is likely to guarantee the query image actually resembles to the first closest image from the library (for k = 1). However, by raising the value of k to 3 and 5, the library images tend to be dissimilar to the query image, which results in a lower performance (in particular, the SEN). The rationale behind such accuracy decrease can be referred to the limited number of library images. In other terms, for every indoor scenery, there are few representative images within the library. Increasing such number would certainly promote a better correlation between the k considered library images and uplift the probability of having objects in common and hence boost the fusion process but at the cost of a larger processing time. On the other hand, such phenomenon is not observed with the SSCS strategy since similarity computation is performed not in the image domain but in the semantic one. Moreover, it tends to be more balanced between the SEN and the SPE, which is not the case with the former strategy. The result differences between the two datasets can be explained by the fact that the structure and the quantity of the objects composing their images, in addition to the physical dimensions of the two buildings, are different. For both datasets, and by averaging the SEN and the SPE, the best outcomes were obtained for k = 3 using the semantic similarity, and for k = 1 regarding the Euclidean distance. In general, the SSCS, in spite of the small-size library, behaves better than the EDCS given that it performs the multilabeling process more efficiently. This is because image similarity assessment in the semantic domain appears more straightforward to infer than in the image domain which is more sensitive to image acquisition condition issues. As for the behavior of the GP regressors, it can be drawn that the obtained results are very satisfactory despite that only few training images in both datasets are used.

To analyze in more detail the obtained results, we also provide the recognition accuracies of each object. They are summarized in Tables 3 and 4. In particular, the objects range from 1 to 15 pointing, respectively, to the following categories.

Dataset 1: External Window, Board, Table, External Door, Stair Door, Access Control Reader, Office, Pillar, Display Screen, People, ATM, Chairs, Bins, Internal Door, and Elevator.

Dataset 2: Stairs, Heater, Corridor, Board, Laboratories, Bins, Office, People, Pillar, Elevator, Reception, Chair, Self-Service, External Door, and Display screen.

Table-3. Per-Class, Sen, and Speaccuracies achieved on Dataset 1 by: Edcs method (k = 1 AND 1/10 RATIO), SSCS Method (k = 3 AND 1/2 RATIO), and Surf-Based method.

Objects	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	SEN	SPE
EDCS	77.77	61.11	93.05	75	77.77	90.27	75	66.66	77.77	95.83	91.66	65.27	66.66	59.72	87.50	71.53	79.33
SSCS	58.33	59.72	91.66	69.44	63.88	90.27	44.44	54.16	90.27	95.83	87.50	38.88	72.72	73.61	88.88	80.52	69.74
SURF	83.33	81.94	100	93.05	90.27	95.83	79.16	90.27	93.05	95.83	93.05	88.88	93.05	79.16	100	71.16	100

Table-4. Per-Class, Sen, and Speaccuracies achieved on Dataset 2 by: Edcs method(k = 1 AND 1/2 RATIO), SSCS Method (k = 3 AND 1/2 RATIO), and Surf-Based method.

Objects	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	SEN	SPE
EDCS	60	74.28	68.57	50	65.71	67.14	78.57	82.85	75.71	87.14	78.57	91.42	77.14	77.14	81.42	70	90.12
SSCS	45.71	62.85	64.28	45.71	64.28	82.85	67.14	88.57	80	91.42	81.42	95.71	90	90	85.71	70.90	82.65
SURF	98.57	94.28	85.71	62.85	62.85	87.14	91.42	90	97.14	90	90	98.57	100	98.57	95.71	77.72	100

The per-class accuracies vary from 38.88% to 95.83% for the first dataset, while they range from 45.71% to 85.71% for the second one. The poor accuracies obtained for some objects are mainly due to the small number of training images covering them. Augmenting the number of training images is, as mentioned earlier, a possible but not attractive (in our application context) solution to improve the accuracies, since it would impact directly on the processing time.

In relation to the effect of the image resolution, there is no significant change in terms of accuracies. However, there is a progressive decrease (while dropping the resolution ratio from 1 to 1/10) in terms of average processing time per image, as given in Table-5. It is to mention that the processing time per image is very satisfactory as the experimentations were conducted by means of MATLAB (R2013a), meaning that it could be further reduced if a real-time-oriented programming language were used. Also noteworthy is that, raising the number of the CS learning images could convey a richer representation of the training/test images, but again at the cost of a larger processing time.

Table-5. Average processing time per image (in seconds) for both strategies and for different image resolutions (k = 3).

Ratio	1/10	1/5	1/2	1
Dataset 1 SSCS	1.17	1.22	1.42	2.16
Dataset 1 EDCS	1.08	1.1	1.41	2.44
Dataset 2 SSCS	1.17	1.21	1.53	2.66
Dataset 2 EDCS	1.2	1.23	1.54	2.69

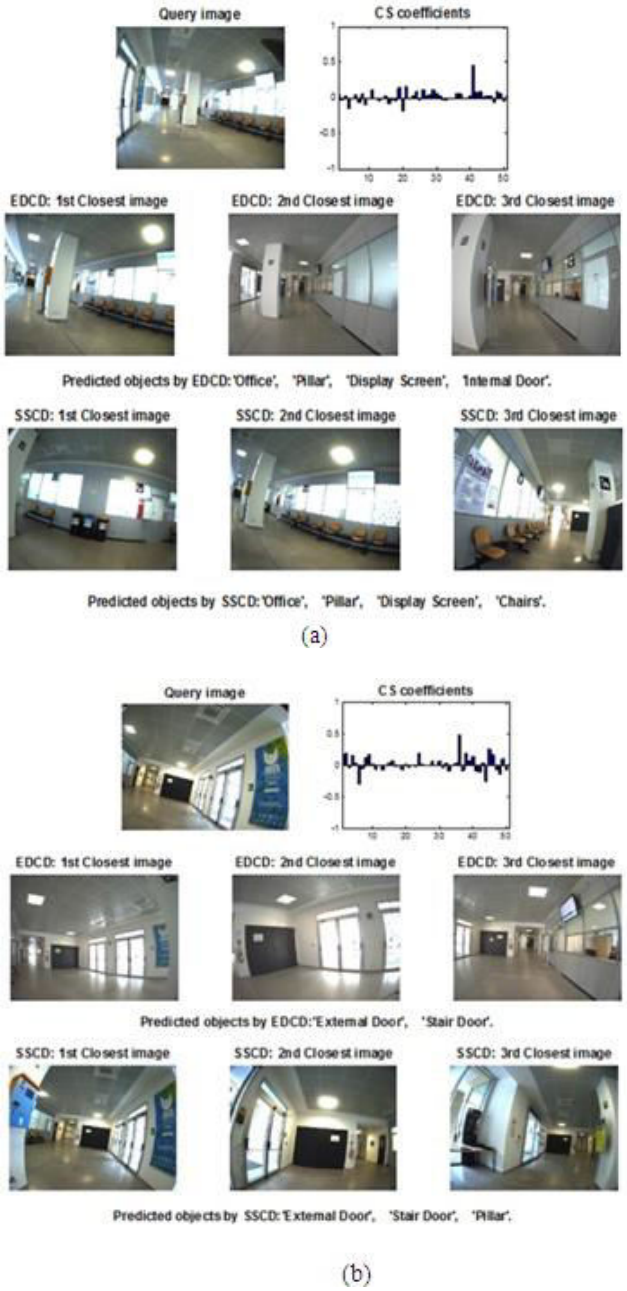

Figure-6. (a) and (b) Two examples of coarse image description conducted on the first dataset. First row: query image and its CS representation. Second row: three most resembling images using EDCS. Third row: three most resembling images using SSCS.

Figures 7 and 8 provide some examples of coarse image description (for $k = 3$) for Datasets 1 and 2, respectively. It can be seen from the examples, considering the EDCS strategy that the first image from the library is apparently similar to the query subject, however, the second and third ones are not necessarily seemingly close to it. By contrast, the SSCS strategy is featured by a more semantic harmony among the identified closest library images.

For the sake of comparison, we run a reference method based on local features, called SURF [41]. This method, well known for its accurate and fast processing capabilities, was originally developed for the detection of single objects. The underlying concept of the SURF-based method is that a number of salient features, denoted keypoints, is extracted from a given test image and then matched to an ensemble of beforehand prepared templates of the object to check which of them is possibly contained in the target image. For our experiments, we have exploited the algorithm available in [42], which incorporates the SURF-based matching method [41] and geometric transformations [43], [44] to determine the bounding box surrounding the object being searched for. Since we are dealing with a multiobject recognition problem, for each class of objects, we cropped its corresponding templates from the training images containing it.

Given a test (query) image, we proceed by checking the presence of all the templates pertaining to all the available classes of objects. If at least one template of a given class is detected in the query image, we consider its class as present in the scene. Otherwise, the object is assumed absent. The training/test images used for the SURF-based method are the same as those used for both EDCS and SSCS techniques. The per-class accuracies as well as the SEN and SPE values achieved by this reference method on both datasets are reported in Tables 3 and 4, respectively.

In overall terms, it can be pointed out that the SURF-based method performs better than EDCS and SSCS. This expected result is motivated by the fact that this method captures local details within the image and uses them to describe the objects, making thus the recognition task potentially more precise than our proposed strategies, which instead deal with the images as single entities. On the other hand, whilst SURF-based method analyzes thoroughly the images, it is comparatively far more computationally demanding. Indeed, while our approach performs over 1 s/image, the SURF-based method requires around 46 s/image on the same machine. Such a processing time cannot be envisioned in our application context. In Figure-9, we provide the outcomes of the SURF-based method for the same test images shown in Figures 7 and 8.

C. Robustness to illumination

To evaluate the SEN of the proposed strategies to illumination problems, we considered a scenario in which the training and test images are acquired in two separate daytimes. In greater detail, the dataset is composed of 51 training images captured in the morning and 79 test images acquired in the evening.

Table-6. Per-Class, SEN, and SPE accuracies obtained for the separate daytimes scenario, with EDCS (k = 1 AND 1/10 RATIO) and SSCS (k = 3 AND 1/2 RATIO).

Objects	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	SEN	SPE
EDCS	87.50	54.16	94.44	66.66	51.38	90.27	44.44	80.55	80.55	95.83	87.50	43.05	54.16	48.61	68.05	52.90	84.34
SSCS	86.11	52.77	94.44	48.61	48.61	91.66	51.38	44.44	45.83	45.83	65.27	51.38	52.77	52.77	88.88	72.25	68.00

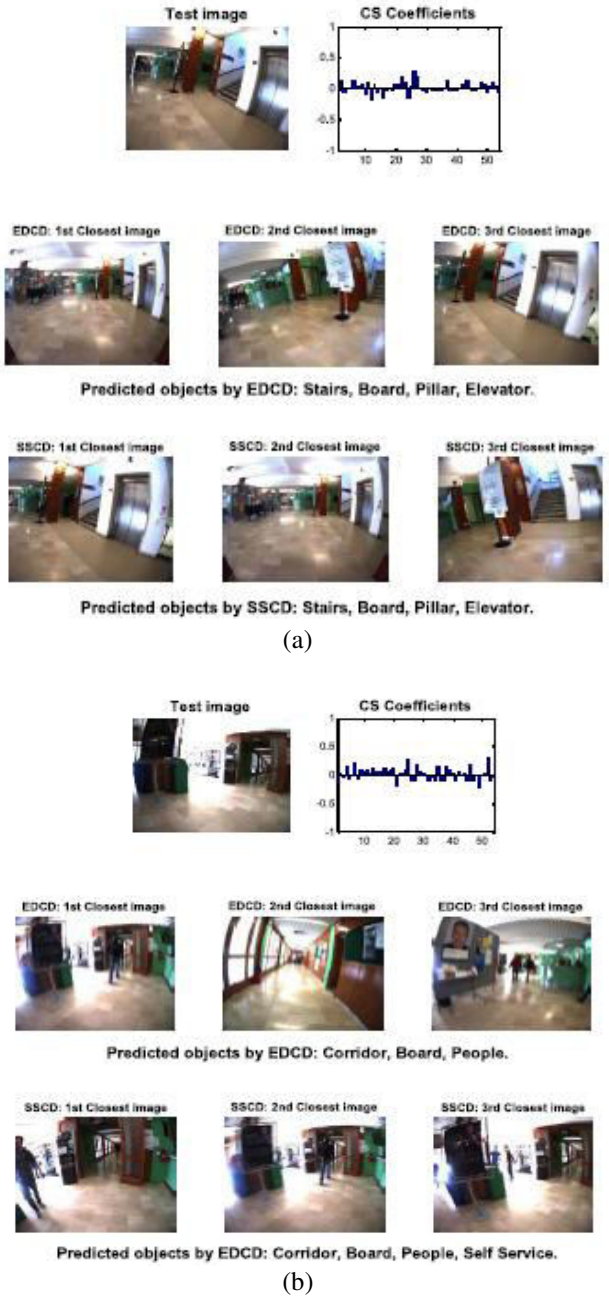

Figure-7. (a) and (b) Two examples of coarse image description corresponding to the second dataset. First row: query image and its CS representation. Second row: three most resembling images using EDCS. Third row: three most resembling images using SSCS.

In this experiment, the values of k (number of multilabeling images) and the image resolution ratio refer to the ones which yielded the best accuracies in Table I (i.e., EDCS: k = 1 and a 1/10 ratio, SSCS: k = 3 and a 1/2 ratio). The related results are summarized in Table VI, where both per-class accuracies as well as SEN and SPE metrics are listed. From these results, it can be noticed that the SEN of the EDCS strategy is about 25% less than that of SSCS. This supports that comparing images in the semantic domain is more effective than a direct distance computation in a given image domain representation (in our case in the CS coefficient domain). In a real-implementation scenario, we recommend to construct the set of training images in such a way that it conveys maximum representativeness to cover all the predefined objects and various possible image acquisition conditions (e.g., changes of illumination).

CONCLUSIONS

A common way of tackling the issue of object recognition under a blind rehabilitation prospect is the reliance on detecting one specific kind of objects. As to relay more information on the scene under analysis, broadening the emphasis to multiple objects becomes necessary but raises implementation issues due to the very tight time constraint. To this end, this paper introduces a novel multiobject detection approach for indoor scenes through coarse image description, which is fulfilled by multilabeling an image acquired by a camera mounted on the user. Coarse image description was considered with the aim to enhance the perception and the comprehension of a blind individual to his/her nearby objects in an indoor environment.

Figure-8. Results of SURF-based method achieved on the examples in Figures 7 and 8. First row for dataset 1 and second row for dataset 2. Detected objects are as follows. Top left: External door, Office, Pillar, Display screen, and Chairs. Top right: Board, External door, and Stair door. Bottom left: Stairs, Board, Pillar, and Elevator. Bottom right: Corridor, Board, Laboratories, and Self-Service.

The idea is to make use of an offline prepared library consisting of different images captured from different points distributed all over the considered indoor environment. Image representation was dealt with through a CS-based technique to guarantee compactness, and thus short image analysis time. The query image is coarsely described by fusing a given number of most similar images from the library. In this context, the similarity concept was carried out by proposing two strategies, a basic Euclidean distance strategy and a semantic-based similarity. The latter one is preferred over the former one since it gauges the resemblance between images on the basis of their semantic content and not on their spectral appearances.

Pros and cons of the present methodology can be summarized as follows. The two main advantages are: 1) its capability to detect simultaneously numerous simple as well as complex objects and 2) contained processing time. This makes it particularly suited for (quasi) real-time applications like assistive technologies for blind people. On the other side, its main drawback is related to the design of the set of training images, which needs a particular care. Indeed, a small training set favors the processing time but at the expense of the recognition accuracy, and vice versa. In general, we recommend that it conveys maximum representativeness to cover all the predefined objects and various possible image acquisition conditions, keeping into account that additional gains in processing times can be obtained by reducing image resolution with very limited impact on accuracy.

As for future concerns, we think that our framework could be further improved by considering the following strategies:

1) Developing an image rejection phase on the library images prior to proceeding with the multilabeling process to discard outlier images in the fusion stage.

2) Applying a weighted sum on the k binary descriptors, while performing the fusion could also be an interesting way to follow but raises the problem of the estimation of the best weight values.

REFERENCES

[1] D. L. Donoho. 2006. Compressed sensing.IEEE Trans. Inf. Theory. 52(4): 1289-1306.

[2] E. J. Candès, J. Romberg and T. Tao. 2006. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory. 52(2): 489-509.

[3] L. Dandona and R. Dandona. 2006. Revision of visual impairment definitions in the international statistical classification of diseases.BMC Med.4: 1-7.

[4] M. da Silva Cascalheira, P. Pinho, D. Teixeira and N. B. de Carvalho. 2012. Indoor guidance system for the blind and the visually impaired.IET Microw. Antennas, Propag. 6(10): 1149-1157.

[5] I. Ulrich and J. Borenstein. 2001. The GuideCane-Applying mobile robot technologies to assist the visually impaired. IEEE Trans. Syst., Man, Cybern. A, Syst., Humans. 31(2): 131-136.

[6] S. Shoval, J. Borenstein, and Y. Koren. 1998. Auditory guidance with the Navbelt-A computerized travel aid for the blind. IEEE Trans. Syst., Man, Cybern. C, Appl. Rev. 28(3): 459-467.

[7] M. R. Strakowski, B. B. Kosmowski, R. Kowalik, and P. Wierzba. 2013. An ultrasonic obstacle detector based on phase beamforming principles. IEEE Sensors J. 6(1): 179-186.

[8] S. Pundlik, M. Tomasi and G. Luo. 2013. Collision detection for visually impaired from a body-mounted camera. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW). pp. 41-47.

[9] V. Pradeep, G. Medioni and J. Weiland. 2010. Robot vision for the visually impaired. In:Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW). pp. 15-22.

[10] M. Radvanyi, B. Varga and K. Karacs. 2010. Advanced crosswalk detection for the bionic eyeglass. In:Proc. 12th Int. Workshop Cellular Nanoscale Netw. Appl. (CNNA). pp. 1-5.

[11] G. Balakrishnan, G. Sainarayanan, R. Nagarajan and S. Yaacob. 2004. Stereopsis method for visually impaired to identify obstacles based on distance. In: Proceeding of IEEE 1st Symp. Multi-Agent Secur. Survivability. pp. 580-583.

[12] F. M. Hasanuzzaman, X. Yang, and Y. Tian. 2012. Robust and effective component-based banknote recognition for the blind. IEEE Trans. Syst., Man, Cybern. C, Appl. Rev. 42(6): 1021-1030.

[13] D. López-de-Ipiña, T. Lorido and U. López. 2011. BlindShopping: Enabling accessible shopping for visually impaired people through mobile technologies. In:Proc. 9th Int. Conf. Toward Useful Services Elderly People Disabilities. pp. 266-270.

[14] E. Tekin and J. M. Coughlan. 2009. An algorithm enabling blind users to find and read barcodes. In:Proc. Workshop Appl. Comput. Vis. (WACV). pp. 1-8.

[15] H. Pan, C. Yi and Y. Tian. 2013. A primary travelling assistant system of bus detection and recognition for visually impaired people. In: Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW). pp. 1-6.

[16] T. J. J. Tang, W. L. D. Lui and W. H. Li. 2012. Plane-based detection of staircases using inverse depth. In: Proc. ACRA. pp. 1-10.

[17] X. Chen and A. L. Yuille. 2004. Detecting and reading text in natural scenes. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR). 2: II-366-II-373.

[18] X. Yang and Y. Tian. 2010. Robust door detection in unfamiliar environments by combining edge and corner features. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW). pp. 57-64.

[19] S. Wang and Y. Tian. 2012. Camera-based signage detection and recognition for blind persons. In: Proc. 13th Int. Conf. Comput. Helping People Special Needs. pp. 17-24.

[20] J. Shotton, J. Winn, C. Rother and A. Criminisi. 2009. TextonBoost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. Int. J. Comput. Vis. 81(1): 2-23.

[21] Razavi, J. Gall, and L. Van Gool. 2011. Scalable multi-class object detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 1505-1512.

[22] T. Deselaers, D. Keysers, R. Paredes, E. Vidal, and H. Ney. 2003. Local representations for multi-object recognition. In: Pattern Recognition. Berlin, Germany: Springer-Verlag. pp. 305-312.

[23] K. Mikolajczyk, B. Leibe and B. Schiele. 2006. Multiple object class detection with a generative model. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 26-36.

[24] C. Pantofaru, C. Schmid and M. Hebert. 2008. Object recognition by integrating multiple image segmentations. In: Proc. 10th Eur. Conf. Comput. Vis. (ECCV). pp. 481-494.

[25] M. Aharon, M. Elad, and A. Bruckstein. 2006. K-SVD: An algorithm for designing over completes dictionaries for sparse representation. IEEE Trans. Signal Process. 54(11): 4311-4322.

[26] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma. 2009. Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. 31(2): 210-227.

[27] V. M. Patel and R. Chellappa. 2011. Sparse representations, compressive sensing and dictionaries for pattern recognition. In: Proc. 1st Asian Conf. Pattern Recognit. (ACPR). pp. 325-329.

[28] A. Morelli Andrés, S. Padovani, M. Tepper and J. Jacobo-Berlles. 2014. Face recognition on partially occluded images using compressed sensing. Pattern Recognit. Lett. 36: 235-242.

[29] J. Yang, J. Wright, T. S. Huang and Y. Ma. 2010. Image super-resolution via sparse representation. IEEE Trans. Image Process. 19(11): 2861-2873.

[30] S. Rao, R. Tron, R. Vidal, and Y. Ma. 2010. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. IEEE Trans. Pattern Anal. Mach. Intell. 32(10): 1832-1845.

[31] J. Mairal, M. Elad and G. Sapiro. 2008. Sparse representation for color image restoration. IEEE Trans. Image Process. 17(1): 53-69.

[32] B. Shen, W. Hu, Y. Zhang and Y.-J. Zhang. 2008. Image inpainting via sparse representation. In: Proc. IEEE ICASSP. pp. 697-700.

[33] L. Lorenzi, F. Melgani and G. Mercier. 2013. Missing-area reconstruction in multispectral images under a compressive sensing perspective. IEEE Trans. Geosci. Remote Sens. 51(7): 3998-4008.

[34] A. Quattoni, M. Collins and T. Darrell. 2008. Transfer learning for image classification with sparse prototype representations. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 1-8.

[35] J. Wright, Y. Ma, J.Mairal, G. Sapiro, T. S. Huang and S. Yan. 2010. Sparse representation for computer vision and pattern recognition. Proc. IEEE. 98(6): 1031-1044.

[36] E. J. Candès and T. Tao. 2005. Decoding by linear programming.IEEE Trans. Inf. Theory. 51(12): 4203-4215.

[37] D. L. Donoho, Y. Tsaig, I. Drori and J.-L. Starck. 2012. Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. IEEE Trans. Inf. Theory. 58(2): 1094-1121.

[38] C. K. I. Williams and D. Barber. 1998. Bayesian classification with Gaussian processes. IEEE Trans. Pattern Anal. Mach. Intell. 20(12): 1342-1351.

[39] C. E. Rasmussen and C. K. I. Williams. 2006. Gaussian Processes for Machine Learning. Cambridge, MA, USA: MIT Press.

[40] Y. Bazi and F. Melgani. 2010. Gaussian process approach to remote sensing image classification. IEEE Trans. Geosci. Remote Sens. 48(1): 186-197.

[41] H. Bay, T. Tuytelaars and L. Van Gool. 2006. SURF: Speeded up robust features. In: Proc. 9th Eur. Conf. Comput. Vis. (ECCV). pp. 404-417.

[42] SURF-Based Object Detection. [Online]. Available: http://www.mathworks.it/it/help/vision/examples/object-detection-in-a-clutteredscene- using-point-feature-matching.html, accessed June 25.

[43] R. Hartley and A. Zisserman. 2003. Multiple View Geometry in Computer Vision. Cambridge, U.K.: Cambridge Univ. Press.

[44] P. H. S. Torr and A. Zisserman. 2000. MLESAC: A new robust estimator with application to estimating image geometry. Comput. Vis. Image Understand. 78(1): 138-156.

[45] SparseLab. [Online]. Available: http://sparselab.stanford.edu, accessed Feb. 1.

[46] C. E. Rasmussen and K. I. Williams. 2014. Gaussian Process Soft-Ware [Online]. Available: http://www.Gaussianprocess.org/gpml/code/matlab/doc/, accessed June 25.