Part-Based Room Categorization for Household Service Robots

Peter Uršič[1*] and Rok Mandeljc[1*] and Aleš Leonardis[2] and Matej Kristan[1]

***Abstract*— A service robot that operates in a previously-unseen home environment should be able to recognize the functionality of the rooms it visits, such as a living room, a bathroom, etc. We present a novel part-based model and an approach for room categorization using data obtained from a visual sensor. Images are represented with sets of unordered parts that are obtained by object-agnostic region proposals, and encoded using state-of-the-art image descriptor extractor — a convolutional neural network (CNN). An approach is proposed that learns category-specific discriminative parts for the part-based model. The proposed approach was compared to the state-of-the-art CNN trained specifically for place recognition. Experimental results show that the proposed approach outperforms the holistic CNN by being robust to image degradation, such as occlusions, modifications of image scaling, and aspect changes. In addition, we report non-negligible annotation errors and image duplicates in a popular dataset for place categorization and discuss annotation ambiguities.**

I. Introduction

Over the last decade, a significant effort has been invested in development of service robots, with the aim of constructing machines that would, ideally, be capable of autonomously performing services for human well-being. A crucial capability of the service robots coexisting with humans in their homes is recognition of household space categories. Several space categorization approaches have been proposed recently that rely on laser-range sensors [1], [2], [3], and an increasingly large body of literature focuses on image-based recognition (e.g., [4], [5], [6]). A de-facto standard approach for image-based recognition encodes the image as a compact feature vector and classifies it using a pre-trained classifier. In most cases, the image is encoded by a bag-of-words-like model (e.g., [7], [8]), formed from low-level [9] or mid-level [5] descriptors, and recently the availability of large image datasets [10], [11] has fostered the renaissance of convolutional neural networks (CNNs) [12], [13], which by far outperform other holistic image descriptors on the task of place recognition [11].

The boost in performance of the holistic models for place recognition [11] can be largely attributed to the excellent discriminative properties of the CNN features. Nevertheless, their formulation inherently assumes that the images of places do not contain occluders, and that they are captured by image sensors with similar width-to-height ratios. Such assumptions are unrealistic for applications in service robots.

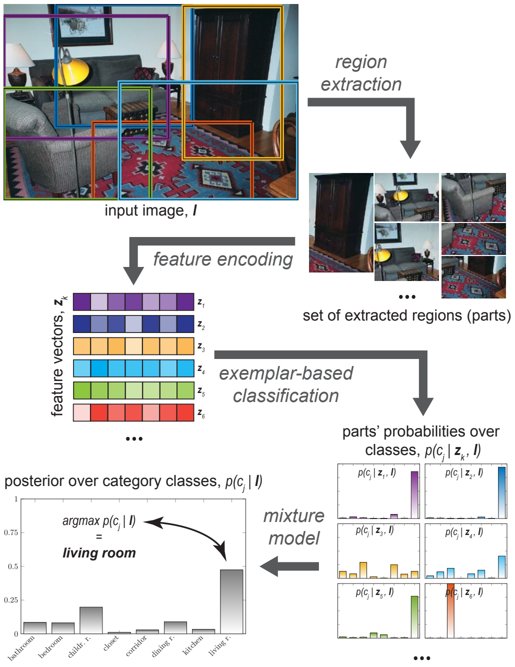


Fig. 1. Outline of the proposed part-based room categorization approach.

In a typical household, the image of the observed room may be partially occluded by a wall or a person standing in front of the robot. The room composition may also vary significantly across the images of the same category. Moreover, the assumption of similar width-to-height ratios (aspect) prevents an offline training of a general room categorization system for its later application to arbitrary-sized sensors on a particular robot.

The occlusion, aspect, and composition variations are conceptually better addressed by the part-based models [14], [15], [16]. These models decompose the image into segments [14] or objects [15], and perform classification. Nevertheless, several challenges remain: (i) How to efficiently identify image regions that form parts without assuming a fixed image aspect? (ii) How to select parts that generalize well within category and discriminate across different categories? (iii) Based on identified library of discriminative parts, how to construct a classifier that does not require a holistic representations and is robust to occlusion?

A. Our approach and contributions

In this paper, we propose a novel part-based model for room categorization in household service robots. Our first contribution is formulation of the categorization as

[*] Both authors contributed equally.

[1] Authors are with Faculty of Computer and Information Science, University of Ljubljana {peter.ursic, rok.mandeljc, matej.kristan}@fri.uni-lj.si

[2] Aleš Leonardis is with CN-CR Centre, School of Computer Science, University of Birmingham a.Leonardis@cs.bham.ac.uk

a maximum-a-posteriori inference on a mixture model of detected parts (Fig. 1). Rather than sampling parts from a pre-defined grid [14] or resorting to random sampling [6], we apply object-type-agnostic part generation by region proposals [17] that follow the gestalt principle. Such approach generates a large number of region proposals that may correspond to individual objects, visually-coherent compositions of objects, or simply segments of scene with coherent characteristics like color or texture. Each region is encoded by a generally-trained feature descriptor (CNN [11]). Our second contribution is proposing the approach for discriminative construction of part dictionary that is used in our mixture model. The proposed part-based approach is tested on eight categories from the MIT Indoor67 dataset [15] that represent household spaces. All images have been manually inspected to ensure the correctness of labels. We have discovered that 4% of images were originally [15] incorrectly labeled. We have re-labeled those images and will make the labels publicly available. As the MIT Indoor67 dataset is widely used in performance evaluation, e.g., [11], [4], [5], [6], [18], [19], [20], [21], [16], [22], [23], [24], [25], [26], [14], [27], the notion of 4% annotation error will likely contribute to revising the results obtained so far on this dataset. We show that the proposed model outperforms the top-performing approach [11] in robustness to occlusion and aspect change. In addition, we provide qualitative evaluation, offering additional insights into the performance of our approach as well as pointing out ambiguities in the dataset labels.

The remainder of the paper is organized as follows. Section II provides an overview of the related work, our part-based approach is presented in Section III, experimental results are discussed in Section IV, and Section V concludes the paper.

II. Related Work

Most approaches for room categorization apply holistic image descriptors. Typically, the holistic representations are formed in a bag-of-words fashion [7], by forming histograms of different feature occurrences in images. For example, in a robotic setup, a bag-of-words model using SIFTs was used by Ayers *et al.* [28], while Wu *et al.* [29] apply local texture descriptors as visual words. Recently, Dixit *et al.* [26] applied oriented texture curves, and Gemert *et al.* [30] applied proto-concepts as region-based visual words.

Several researchers [31], [24] apply a large number of object detectors, and reason about scene category based on the occurrences of the detected objects. Mesnil *et al.* [25] learn co-occurrences of the object detectors from [24] to form a holistic representation. Such approaches largely depend on the quality of the detectors and representativity of the learned objects for the particular application domain. In [5], authors learn distinctive region detectors, which are in turn used to construct a holistic "bag of parts" descriptor. Parizi *et al.* [6] jointly learn a set of prototype regions and image classifiers by random sampling of prototype regions and forming a holistic image representation by concatenation of the region responses.

To include spatial information in the holistic descriptors, several researchers apply spatial pyramid histograms [23], [20]. Sadovnik *et al.* [32] apply hierarchical grouping of objects in the image. They train individual object detectors and object group detectors, whose responses are then max-pooled in a spatial pyramid to form final image representation. Doersch *et al.* [21] discover visually-coherent patch clusters from an image collection that are maximally discriminative with respect to the labels. Image representation is constructed using a spatial pyramid by taking the max score per detector for each region.

A great boost in performance has been recently achieved by the application of the convolutional neural networks. In [11], for example, the authors achieve state-of-the-art place categorization performance by extracting CNN features from images. Several researchers, e.g., [22], [4] proposed concatenating the CNN features with additional features to increase the performance of the final descriptor. Authors of [19] propose visual words extraction by CNN patches sampled at multiple scales. Dixit *et al.* [27] classify image patches using CNN and apply Fisher Vector embedding for spatial encoding.

To exploit the rich structure that is present in images, part-based models have been considered. Parizi *et al.* [14] proposed reconfigurable models that split the images into pre-defined grid and classify each cell into a predefined class. The inferred spatial configuration is used as final descriptor. Their approach highly depends on selecting the specific grid decomposition and manually selected number of part classes. Quattoni *et al.* [15] form a global image description by GIST descriptor [33] and encode human-annotated regions by spatial pyramids [8]. But only a few parts per image are considered and their position is highly constrained. Pandey *et al.* [16] adapted the deformable parts model (DPM) originally designed for object detection, but used LSVM training [34] to discover a single informative image region for place categorization automatically. In contrast to our approach, their approach is based on a greedy sliding window search, cannot handle aspect changes, and assumes a single informative part per image.

III. Part-based Room Categorization

In our approach, an image $\boldsymbol{I}$ is represented by an automatically-extracted unordered set of image regions (parts) $\{\boldsymbol{z}_k\}_{k=1:N_P}$, where $\boldsymbol{z}_k$ is a feature vector corresponding to the k-th region (see Fig. 1). The region extraction approach is discussed in Section III-A. The extracted regions are used in the learning stage to construct non-parametric part-based models (Section III-C), and in classification stage to infer the most likely image category (Section III-B).

A. Object-agnostic part region extraction

The obvious choices of object-agnostic region extractors are greedy enumeration of all possible regions of different dimensions and scales, random sampling, or region extraction at pre-defined locations. But none of these approaches guarantee repetitive region detection on salient structures that occur across the images of the same category. Recently, a variety of methods for generating category-independent region proposals that likely contain highly informative structures have been presented [17], [35], [36]. We have chosen the gestalt-principle-inspired selective search [17], due to its great success in the recent approach for object detection [13], but other region proposals could be considered as well.

Application of the selective search [17] to the image yields a set of regions that need to be encoded in a feature space. The choice of features has a great impact on the overall system performance. Conventionally, hand-crafted features have been used, such as [9], [37], [7], [38], and recently, the convolutional neural networks (CNN) [12], [13], [11] have been shown to significantly outperform these methods, and perform exceptionally well as generic feature extractors [18]. For these reasons we apply the hybrid-CNN from [11] as a region feature extractor.

B. Part-based categorization model

Given the set of extracted parts $\{\boldsymbol{z}_k\}_{k=1:N_P}$, the image $\boldsymbol{I}$ is classified as category c_{opt} by maximizing the posterior over the category classes, i.e., $c_{\text{opt}} = \arg\max_{c_j} p(c_j|\boldsymbol{I})$. The posterior is defined by a part-based mixture model,

$$p(c_j|\boldsymbol{I}) = \sum_{k=1:N_P} p(c_j|\boldsymbol{z}_k, \boldsymbol{I}) p(\boldsymbol{z}_k|\boldsymbol{I}), \tag{1}$$

where $p(c_j|\boldsymbol{z}_k, \boldsymbol{I})$ is the probability of category c_j given the observation $\boldsymbol{z}_k$, and $p(\boldsymbol{z}_k|\boldsymbol{I})$ is the prior on observation $\boldsymbol{z}_k$. The former probability is modeled by a logistic function

$$p(c_j|\boldsymbol{z}_k, \boldsymbol{I}) = \frac{\gamma}{1 + \exp(a f_j(\boldsymbol{z}_k) + b)}, \tag{2}$$

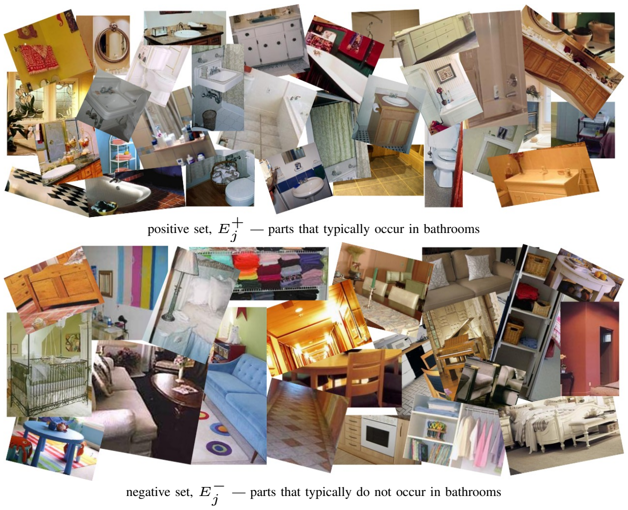
where $f_j(\boldsymbol{z}_k)$ is the output of an exemplar-based classifier for j-th category, (a, b) are learned parameters, and γ is a normalization constant ensuring that the posterior (2) sums to one over all c_j. The prior should reflect the region informativeness for classification; the observations $\boldsymbol{z}_k$ that contain a peaked likelihood across the alternative categories (2) should contribute more to the posterior (1) than those with poorly expressed peaks. Thus we define the prior as the peak value of the likelihood, i.e.,

$$p(\boldsymbol{z}_k|\boldsymbol{I}) = \max_j p(c_j|\boldsymbol{z}_k, \boldsymbol{I}). \tag{3}$$

The exemplar-based classifier $f_j(\boldsymbol{z}_k)$ in (2) is defined by the set of positive E_j^+ and negative E_j^- exemplars, i.e., $E_j = \{E_j^+, E_j^-\}$. Given a definition of the distance function between an exemplar $\mathbf{x}_i$ and the observation $\boldsymbol{z}_k$, $K(\mathbf{x}_i, \mathbf{z}_k)$, the classification function can be defined in a standard form as

$$f_j(\boldsymbol{z}_k) = \sum_{i=1:|E_j|} \alpha_i y_i K(\mathbf{x}_i, \mathbf{z}_k), \tag{4}$$

where $y_i \in \{-1, 1\}$ indicates the negative or positive exemplar and α_i are learned exemplar weights.



positive set, E_j^+ — parts that typically occur in bathrooms

negative set, E_j^- — parts that typically do not occur in bathrooms

Fig. 2. Examples of collected exemplar parts, E_j, for the *bathroom* category.

C. Learning the category-specific exemplars

The efficiency of the probabilistic model (1) largely depends on the expressiveness of the exemplar parts learned for each category. This section formalizes the approach for exemplar learning.

Consider learning the dictionary of exemplars E_j for a category c_j. A set of features $\{\boldsymbol{x}_i\}_{i=1}^N \in X$ is extracted from a set of training images as described in Section III-A. The set is ordered such that the $C_j^+ = \{\boldsymbol{x}_i\}_{i=1}^{N^+}$ correspond to the parts from images of category c_j, and $C_j^- = \{\boldsymbol{x}_i\}_{i=N^+ +1}^N$ correspond to the images from the other categories. Assume a transformation $\phi : X \to X'$, such that $\{\phi(\boldsymbol{x}_i)\}_{i=1}^N \in X'$ are linearly separable with respect to C_j^+ and C_j^-, and $||\phi(\boldsymbol{x}_i)|| = 1$ for $i \in \{1, 2, \ldots, N\}$. Then, the dot product $\langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_l) \rangle$ defines a similarity measure among the elements of X', with the value 1 corresponding to the maximum similarity. The goal is to find non-empty subsets of the exemplars $E_j^+ \subset C_j^+$ and $E_j^- \subset C_j^-$, such that $|E_j^+|$ and $|E_j^-|$ are kept as small as possible, and that for each $\boldsymbol{x}_i$, there exists an exemplar $\boldsymbol{e} \in E_j^u$, for which

$$\langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{e}) \rangle < \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_l) \rangle, \tag{5}$$

for all $\boldsymbol{x}_l \in C_j^{\{+,-\}\backslash u}$. The inequality in Eq. (5) prefers discriminative exemplars. Each $\boldsymbol{x}_i$ is assigned a corresponding variable β_i, where $|\beta_i| \leq B$, which determines the suitability of $\boldsymbol{x}_i$ for being an exemplar, with β_i close to zero indicating small suitability. Without the loss of generality, we define that $\beta_i \to B$ implies that $\boldsymbol{x}_i$ is well-suited for an exemplar of C_j^+, whereas $\beta_i \to -B$ implies that it is well-suited for an exemplar of C_j^-. For ϕ that would sufficiently separate the two categories, a high similarity of two elements $\phi(\boldsymbol{x}_i)$ and $\phi(\boldsymbol{x}_l)$, i.e., $\langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_l) \rangle$, being close to one implies that it is highly probable that $\boldsymbol{x}_i$ and $\boldsymbol{x}_l$ belong to the same category. To induce the sparsity of our solution, we require from such elements that either none or at most one of them are an exemplar. Therefore in such cases, the terms of the form $\beta_i \beta_l \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_l) \rangle$ should be minimized. On the other hand, when $\phi(\boldsymbol{x}_i)$ and $\phi(\boldsymbol{x}_l)$ are sufficiently dissimilar, their dot product is close to zero, and those terms are close to zero regardless of the β_i and β_l values.

Hence, our goal is to minimize the cost $\Psi(\boldsymbol{\beta}) = \sum_{i,l=1}^{N} \beta_i \beta_l \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_l) \rangle$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)$, with respect to $\boldsymbol{\beta}$, which is equivalent to maximization of $-\Psi(\boldsymbol{\beta})$. To avoid a trivial solution with all $\beta_i = 0$, a constraint on the norm of β is added to the cost function, thus

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}(||\boldsymbol{\beta}||_1 - \lambda \Psi(\boldsymbol{\beta})). \tag{6}$$

To enforce equally balanced sets of positive and negative exemplars, we opt for solving Eq. (6) under the condition

$$\sum_{i=1:n} \beta_i = 0. \tag{7}$$

Defining $\beta_i = \alpha_i y_i$, where

$$y_i = \begin{cases} +1 & \text{for } \boldsymbol{x}_i \in C_j^+, \\ -1 & \text{for } \boldsymbol{x}_i \in C_j^-, \end{cases} \tag{8}$$

and setting $\lambda = \frac{1}{2}$, we see that Eqs. (6) and (7), together with $0 \leq \alpha_i \leq B$, represent a constrained optimization problem that corresponds to the dual problem of a soft-margin SVM optimization, with a kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_l) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_l) \rangle$.

Therefore, the exemplars from our problem formulation, which are indicated by the non-zero valued α_i, correspond to the support vectors of the binary SVM classifier defined by (4). Examples of such exemplars for a single category are illustrated in Fig. 2. In a multi-category setting with N_C categories, we can therefore train N_C one-vs-rest SVM classifiers, which provide us with sets of positive and negative exemplars for each category.

Recall that the model $p(c_j|\boldsymbol{z}_k, \boldsymbol{I})$ defined in (2) converts the output of the discriminative function $f_j(\boldsymbol{z}_k)$ into probabilistic output by a logistic function. The parameters (a, b) of this function are estimated from the training set by applying the Platt's calibration [39].

IV. Experimental Results

The proposed part-based approach from Section III was experimentally compared to a state-of-the-art CNN model, trained on two large combined datasets [11], that forms a holistic CNN-based model of the image and applies a SVM for classification. This approach, which we will refer to in the following as the holistic hybrid-CNN, was chosen because of its top-performing results on the recent places categorization benchmark [11].

Below, we provide implementation details in Section IV-A, describe the experiment dataset in Section IV-B, Section IV-C reports quantitative evaluation, and in Section IV-D qualitative results are discussed.

A. Implementation details

In our region proposal, we remove all regions produced by selective search [17] whose diagonal is smaller than $50\,\%$ of the image diagonal. Boxplots demonstrating the number of extracted parts per image corresponding to various categories are shown in Fig. 3. To remove the influence of features, we apply the same CNN as in [11] for encoding the features in parts extracted from the image. In other words, the transformation ϕ from Eq. (5) corresponds to the FC7 layer response of the used CNN architecture, on top of which a linear SVM is applied.

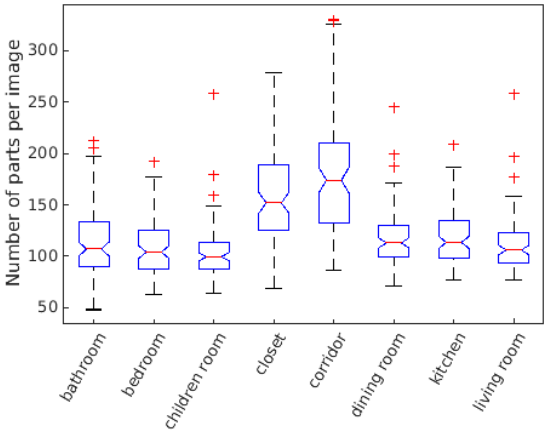


Fig. 3. Boxplots indicating the statistics of the number of extracted parts per image for the household categories.

In particular, we use the Caffe [40] implementation of a deep network called hybrid-CNN, which was pre-trained by [11] on combination of the object-centric ImageNet dataset [10] and the scene-centric Places dataset [11]. In order to compute deep features from a region proposal, the corresponding image patch needs to be first resized to a fixed size of 227×227 pixels, without preserving its aspect ratio. The resized patch is then forward-propagated through five convolutional and two fully-connected layers. Finally, a 4096-dimensional feature vector is obtained by collecting the response from the seventh, fully-connected (FC7), layer of the network. For more details on the network architecture, we refer the reader to [11].

In all experiments, exemplar selection was done using a linear one-vs-rest SVM from LIBLINEAR [41], in combination with zero-mean and unit-variance normalization of features. However, we found that in practice, other multi-class SVM formulations, such as Crammer & Singer multi-class SVM [42], or online (linear) LaRank [43], tend to be also suitable for this task. Furthermore, it is very likely that other multi-class classifiers could be applied as well, such as for example Random Forests [44].

B. The household room dataset

The household room dataset was constructed by collecting the following eight room categories from the MIT Indoor67 dataset [15]: bathroom, bedroom, children room, closet, corridor, dining room, kitchen, and living room. These rooms correspond to the typical categories that a household service robot might encounter in practice.

The authors of MIT Indoor67 dataset [15] provide a pre-defined training/test splits for this dataset, which includes

TABLE I

RESULTS OF CATEGORIZATION USING HOLISTIC HYBRID-CNN [11].

target class label	bathroom	bedroom	child. r.	closet	corridor	dining r.	kitchen	living r.
bathroom	16	0	0	0	0	0	1	0
bedroom	0	17	1	1	0	0	0	2
child. r.	0	1	14	1	0	0	1	1
closet	0	0	0	17	1	0	0	0
corridor	0	0	0	0	21	0	0	0
dining r.	1	0	0	0	0	11	0	1
kitchen	1	0	0	0	0	0	20	0
living r.	1	1	0	0	0	5	1	18

predicted class label (accuracy: 86.45 %)

TABLE II

RESULTS OF CATEGORIZATION USING OUR PART-BASED APPROACH.

target class label	bathroom	bedroom	child. r.	closet	corridor	dining r.	kitchen	living r.
bathroom	17	0	0	0	0	0	0	0
bedroom	0	13	2	1	0	0	0	5
child. r.	0	1	17	0	0	0	0	0
closet	0	0	0	17	1	0	0	0
corridor	0	0	0	0	19	1	1	0
dining r.	2	0	0	0	0	10	1	0
kitchen	0	0	0	0	0	0	21	0
living r.	1	1	1	0	0	4	1	18

predicted class label (accuracy: 85.16 %)

approximately 80 training and 20 testing images per category. For compatibility reasons we apply the same splits in our experiments. To ensure proper labeling, we have manually inspected all the images in the dataset. We have found several (three) images in the training set that were duplicated in the test set. We have removed the duplicates from the training set. We have also found that approximately 4 % of images were incorrectly labeled and have corrected these labels. Nevertheless, there were a few images whose label was ambiguous to us and we left these unchanged (see Section IV-D for example). This revised household room dataset has been made publicly available at the following URL: `http://vision.fe.uni-lj.si/~rokm/icra2016/household_room_dataset`.

C. Quantitative analysis

The performance of the proposed part-based approach was compared to the holistic hybrid-CNN [11] on the household dataset from Section IV-B in two setups. The first setup (baseline experiment) considered the original images from the dataset, and the second setup (robustness study) simulated various modifications of the images, such as occlusion, image scale, and aspect change.

1) Baseline experiment: The results of the baseline experiments are reported as confusion matrices for the hybrid-CNN (Table I) and the proposed part-based approach (Table II). As can be seen, the part-based approach achieves classification accuracy of 85.16 %, which is comparable to that of the hybrid-CNN (86.45 %).

While the performance difference between our part-based approach and the holistic hybrid-CNN is small on the baseline experiment, the difference is most notable on the *children rooms* and *bedrooms* categories. The part-based model outperforms hybrid-CNN on children rooms, but misclassifies bedrooms more often that the hybrid-CNN. As discussed in the qualitative analysis section (Section IV-D), the apparent misclassification can be partly attributed to the ambiguity in room category definition (see Fig. 4 for examples).

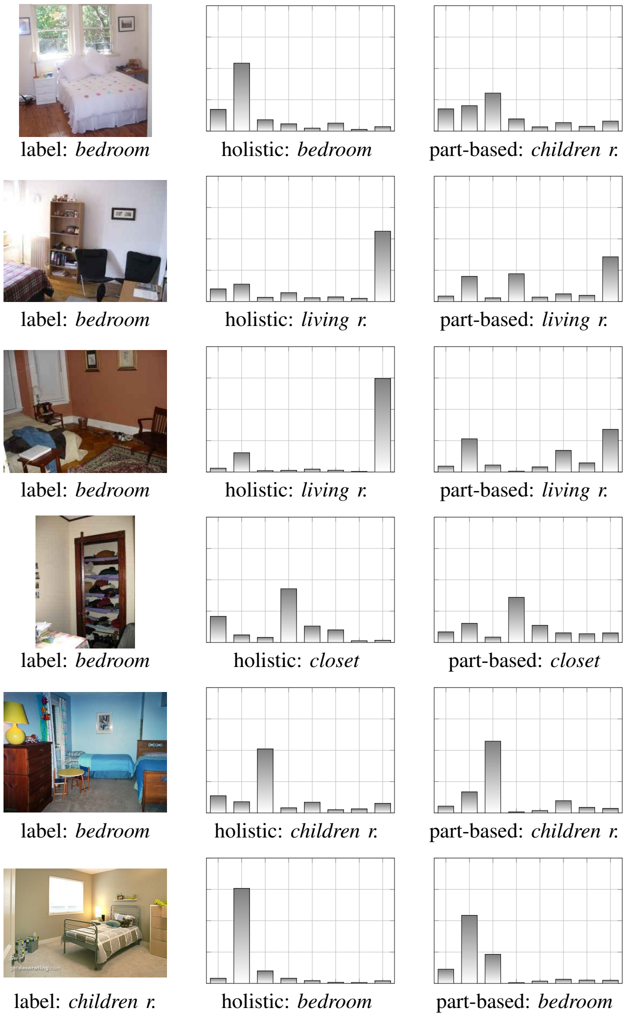


Fig. 4. Examples of test images that are miscategorized primarily due to ambiguous visual content, both by holistic hybrid-CNN [11] and the proposed part-based approach. Bins in the histograms correspond to categories in the following order: bathroom, bedroom, children room, closet, corridor, dining room, kitchen, and living room.

2) Robustness study: In this part of the experiment, we apply various types of image modifications to evaluate the robustness of our approach. In particular, we introduce the following modifications (see Fig. 5 for visual examples):

- *Outside border:* black border is added outside of the image, effectively changing the scale of the original image. The border's width equals a third of the image's larger dimension.
- *Black occluder, right:* the right third of the image is covered by a black vertical stripe.
- *Black occluder, central:* the central third of the image is covered by a black vertical stripe.
- *Person occluder, central:* the central part of the image is covered by an image of a person.

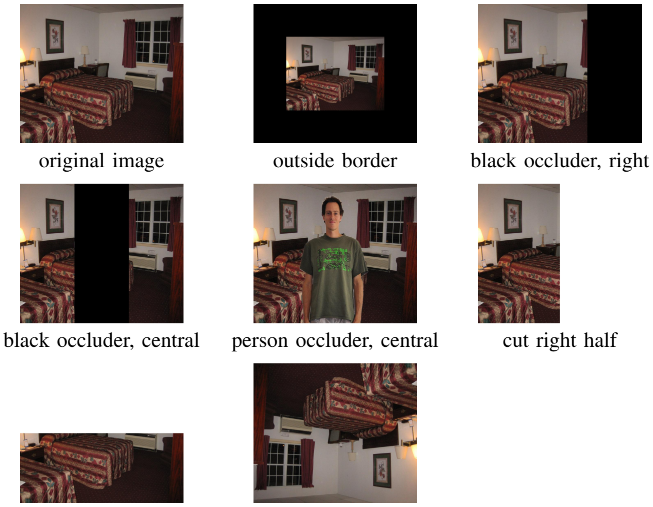

original image
outside border
black occluder, right
black occluder, central
person occluder, central
cut right half
cut top half
upside down

Fig. 5. Examples of input image changes and deformations used in the experiments in Section IV-C.

- *Cut right half:* the right half of the image is removed, reducing the available information, as well as changing the effective aspect ratio.
- *Cut top half:* the top half of the image is removed.
- *Upside down:* the image is rotated for 180°.

The results are summarized in Table III in the form of overall classification accuracy, and show that our part-based approach is consistently less affected by the image modifications that the holistic hybrid-CNN [11]. For example, while the performance of our part-based approach does not change with introduction of the outer border, the performance of the hybrid-CNN drops by more than 20 %. The performance drops in other modifications are less dramatic, but quite apparent. An interesting observation is that either covering the right third or removing the right half of an image causes a similar drop in performance of both approaches, with the part-based approach outperforming the hybrid-CNN by a few percents. The hybrid-CNN is significantly affected by the removal of the top half of the image, whereas the drop in performance is much smaller for our part-based approach. Lastly, neither of approaches were designed to be truly rotation invariant, as shown by the results. Nevertheless, the part-based representation is more flexible, and thus offers improved performance in comparison to the hybrid-CNN. The reason might be that the part-based approach detects parts whose appearance is less rotationally variant than the entire image, which is considered by the holistic model of the hybrid-CNN. In summary, the proposed part-based approach is shown to be much more resilient to the image modifications, as indicated by the lower drop in the classification accuracy compared to the performance on the original images and the average overall performance (Table III).

D. Qualitative analysis

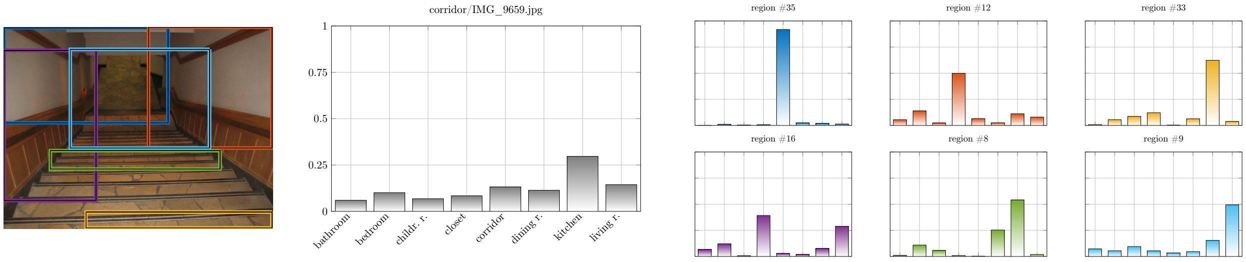
To obtain a deeper insight into operation and failure modes of the proposed part-based approach, we visualize several of the misclassified samples, along with predictions made by both our and the holistic hybrid-CNN models. Figure 4 shows six such examples. The first issue that becomes apparent is that the categories of *bedroom* and *children room* tend to be ambiguous. Even a human observer will have difficulties explaining why the first and the fifth image represent a bedroom (and not a children's room), and conversely, why the last, sixth, image would be categorized as a children room and not a regular bedroom. In the remaining bedroom images, the most prominent feature — bed, is missing from the picture, hence both approaches declare them to be living rooms, or in other instance, a closet. Indeed, the fourth from the top image is labeled as bedroom, but almost entire image is dominated by the closet. Note that in these cases, the proposed part-based approach tends to give less confident predictions than the holistic hybrid-CNN one, which could serve as a measure of the ambiguity with respect to the room's category or function. This is because the part-based approach correctly detects that different parts of the image correspond to different categories. For example, the colorful dots on the bed in the first image of Figure 4 are indicator of children's room, while chairs in the second image indicate a living room.

TABLE III
THE OVERALL CLASSIFICATION ACCURACY FOR EXPERIMENTS ON ORIGINAL AND DISTORTED IMAGES, USING THE HYBRID-CNN [11] AND THE PROPOSED PART-BASED APPROACH.

experiment	**holistic hybrid-CNN [11]**	**part-based approach**
original images	86.45 %	85.16 %
outside border	62.58 %	85.16 %
black occluder, right	78.71 %	80.00 %
black occluder, central	61.94 %	69.68 %
person occluder, central	59.35 %	68.39 %
cut right half	62.58 %	64.52 %
cut top half	52.26 %	68.39 %
upside down	52.26 %	59.35 %
average	64.52 %	72.58 %

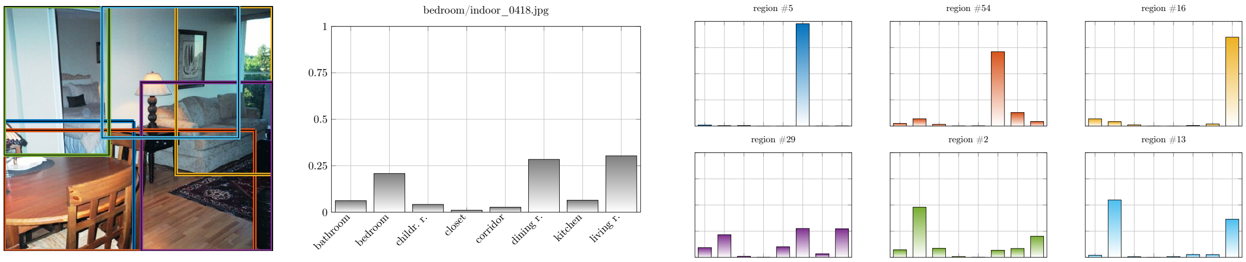
The use of parts allows the proposed approach to capture details at several levels. This can be both advantageous and disadvantageous, as illustrated by Figs. 6 and 7, respectively. Both figures show typical parts found in an input image, their individual category probability predictions, as well as the posterior over room categories. In Fig. 6, a corridor was incorrectly categorized as a kitchen. The regions found in the image tend to capture texture that is not representative for corridors. In particular, a large number of region proposals are obtained for the individual stairs, whose local texture appears to be close to that found in kitchens. Hence, the final posterior is skewed in favor of this category, leading to incorrect categorization.

Figure 7 illustrates the less apparent advantages of part-based representation that are not reflected in the strict single-category prediction setup used in the evaluation. In particular, almost equal posterior probability is obtained for living and dining room, followed by the bedroom — all three room



input image (103 regions) final categorization: *kitchen* individual region categorizations

Fig. 6. An illustration of a failure mode that causes a corridor to be incorrectly categorized as a kitchen.



input image (70 regions) final categorization: *living r.* individual region categorizations

Fig. 7. The advantage of part-based model: prediction when several category types are present in the scene.

categories can in fact be observed in the scene. Even more, due to the localization of relevant parts in the image, the proposed part-based model could be extended to predict functions of certain areas in the room, as opposed to the overall function, which is often ambiguous. Such representation could therefore be used for focus of attention and visual servoing of service robots to relevant parts of the room.

V. Conclusion and Future Work

A part-based approach for visual room categorization appropriate for household service robots has been proposed. Images are represented with sets of unordered parts, which are obtained by extraction of high quality image regions using a region proposal algorithm, and by propagating those regions through several layers of a CNN. An approach is proposed that learns category-specific discriminative parts, and a part-based model is proposed for image categorization. The proposed approach was compared to the state-of-the-art convolutional neural network (CNN) trained specifically for place recognition. Experimental results show that the proposed approach outperforms the holistic CNN by being robust to image degradations, modifications of image scaling, and aspect changes. In addition to the quantitative evaluation, a qualitative evaluation was performed. The results indicate a significant potential of the part-based approach to identify regions in images that actually correspond to different functional categories, thus providing a richer categorization output than the holistic model.

In the process of dataset construction, we have collected and manually verified the household categories from the MIT Indoor67 dataset [15]. We have found that approximately four percent of images were mislabeled, and that the dataset contained duplicated images. We have corrected these errors, and made the revised subset publicly available. The notion of the duplicates and wrong labels is likely to have a significant impact, since a significant number of scene recognition approaches have been evaluated on this dataset.

The spatial distribution of parts corresponding to certain categories could be used for fine-grained image categorization. Robots might acquire images in which more than a single room is visible at a time, moreover, different areas of the same room are often designed for different purposes, e.g., a dining room and a living room can share the same room. If parts corresponding to certain categories are consistently positioned only in particular areas of an image, such areas should be categorized separately. These will be the topics of our future work.

Acknowledgment

This work was supported by the applied project L2-6765 by Slovenian Research Agency and a research programme P2-0214.

References

[1] P. Ursic, M. Kristan, D. Skocaj, and A. Leonardis, “Room classification using a hierarchical representation of space,” *IROS*, 2012, pp. 1371–1378.

[2] O. Mozos, C. Stachniss, and W. Burgard, “Supervised learning of places from range data using adaboost,” *ICRA*, 2005, pp. 1730–1735.

[3] S. Friedman, H. Pasula, and D. Fox, “Voronoi random fields: Extracting the topological structure of indoor environments via place labeling,” *IJCAI*, 2007.

[4] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, “Multi-scale orderless pooling of deep convolutional activation features,” *ECCV*, 2014, pp. 392–407.

[5] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman, “Blocks that shout: Distinctive parts for scene classification,” *CVPR*, 2013.

[6] S. N. Parizi, A. Vedaldi, A. Zisserman, and P. Felzenszwalb, “Automatic discovery and optimization of parts for image classification,” *ICLR*, 2015.

[7] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[8] J. Sivic and A. Zisserman, "Video google: Efficient visual search of videos," *Toward Category-Level Object Recognition*, vol. 4170, 2006, pp. 127–144.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *CVPR*, 2009, pp. 248–255.

[11] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," *NIPS*, 2014, pp. 487–495.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *NIPS*, 2012, pp. 1097–1105.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CVPR*, 2014.

[14] S. Parizi, J. Oberlin, and P. Felzenszwalb, "Reconfigurable models for scene recognition," *CVPR*, 2012, pp. 2775–2782.

[15] A. Quattoni and A. Torralba, "Recognizing indoor scenes," *CVPR*, 2009, pp. 413–420.

[16] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," *ICCV*, 2011, pp. 1307–1314.

[17] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vision*, 2013.

[18] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," *CVPR Workshops*, 2014, pp. 512–519.

[19] Z. Jie and S. Yan, "Robust scene classification with cross-level llc coding on cnn features," *ACCV*, 2015, pp. 376–390.

[20] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," *ECCV*, 2012.

[21] C. Doersch, A. Gupta, and A. A. Efros, "Mid-level visual element discovery as discriminative mode seeking," *NIPS*, 2013, pp. 494–502.

[22] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, and X. Jiang, "Learning discriminative and shareable features for scene classification," *ECCV*, 2014, pp. 552–568.

[23] F. Sadeghi and M. Tappen, "Latent pyramidal regions for recognizing scenes," *ECCV*, 2012, pp. 228–241.

[24] H. Su, L.-J. Li, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," *NIPS*, 2010, pp. 1378–1386.

[25] G. Mesnil, S. Rifai, A. Bordes, X. Glorot, Y. Bengio, and P. Vincent, "Unsupervised learning of semantics of object detections for scene categorization," *Pattern Recognition Applications and Methods*, 2015, vol. 318, pp. 209–224.

[26] R. Margolin, L. Zelnik-Manor, and A. Tal, "Otc: A novel local descriptor for scene classification," *ECCV*, 2014.

[27] M. Dixit, S. Chen, N. D. Gao, Rasiwasia, and N. Vasconcelos, "Scene classification with semantic fisher vectors," *CVPR*, 2015.

[28] B. Ayers and M. Boutell, "Home interior classification using sift keypoint histograms," *CVPR*, 2007.

[29] J. Wu, H. I. Christensen, and J. M. Rehg, "Visual place categorization: Problem, dataset, and algorithm," *IROS*, 2009.

[30] J. van Gemert, J. Geusebroek, C. Veenman, C. Snoek, and A. Smeulders, "Robust scene categorization by learning image statistics in context," *CVPR Workshop*, pp. 105–105, 2006.

[31] J. Yao, S. Fidler, and R. Urtasun, "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation," *CVPR*, 2012, pp. 702–709.

[32] A. Sadovnik and T. Chen, "Hierarchical object groups for scene classification," *ICIP*, 2012, pp. 1881–1884.

[33] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, pp. 145–175, 2001.

[34] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[35] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *TPAMI*, 2012.

[36] J. Carreira and C. Sminchisescu, "Cpmc: Automatic object segmentation using constrained parametric min-cuts," *TPAMI*, vol. 34, no. 7, pp. 1312–1328, 2012.

[37] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *CVPR*, vol. 1, 2005, pp. 886–893.

[38] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," *ECCV*, 2010.

[39] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on platt's probabilistic outputs for support vector machines," *Machine Learning*, vol. 68, no. 3, pp. 267–276, 2007.

[40] Y. Jia, "Caffe: An open source convolutional architecture for fast feature embedding," 2013.

[41] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.

[42] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *J. Mach. Learn. Res.*, vol. 2, pp. 265–292, 3 2002.

[43] A. Bordes, L. Bottou, P. Gallinari, and J. Weston, "Solving multiclass support vector machines with larank," *Proc. of ICML 2007*, 2007, pp. 89–96.

[44] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32.