Robotics and Autonomous Systems 62 (2014) 241–256

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

# Contextual object category recognition for RGB-D scene labeling

Haider Ali[a,*], Faisal Shafait[b], Eirini Giannakidou[c], Athena Vakali[c], Nadia Figueroa[d], Theodoros Varvadoukas[d], Nikolaos Mavridis[d]

[a] Institute of Robotics and Mechatronics (RM), German Aerospace Center (DLR), Oberpfaffenhofen, Germany
[b] The University of Western Australia (UWA), Perth, Australia
[c] Informatics Department, Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece
[d] Department of Engineering, New York University Abu Dhabi (NYUAD), Abu Dhabi, United Arab Emirates

## HIGHLIGHTS

- We use a classifier based on a novel fusion of feature vectors (the VFH-Texton).
- We derive an object-to-object context MRF model based on Flickr label co-occurrence data.
- We investigate the model's parameters' convergence as a function of Flickr's sample size.
- We train the system on the RGB-D Object Dataset and test on the NYU Dataset as well.
- Finally we illustrate an increasing performance through the use of the MRF.

ARTICLE INFO

*Article history:*
Received 3 April 2013
Received in revised form
5 August 2013
Accepted 4 October 2013
Available online 23 October 2013

*Keywords:*
Object recognition
Contextual modeling
RGB-D scenes
Social media
3D scene labeling

## ABSTRACT

Recent advances in computer vision on the one hand, and imaging technologies on the other hand, have opened up a number of interesting possibilities for robust 3D scene labeling. This paper presents contributions in several directions to improve the state-of-the-art in RGB-D scene labeling. First, we present a novel combination of depth and color features to recognize different object categories in isolation. Then, we use a context model that exploits detection results of other objects in the scene to jointly optimize labels of co-occurring objects in the scene. Finally, we investigate the use of social media mining to develop the context model, and provide an investigation of its convergence. We perform thorough experimentation on both the publicly available RGB-D Dataset from the University of Washington as well as on the NYU scene dataset. An analysis of the results shows interesting insights about contextual object category recognition, and its benefits.

## 1. Introduction

Recognizing objects in images and videos is a problem that has been studied in the field of computer vision since its early days. Several decades of research have resulted in object recognition methods that are robust to distortions, view-point changes, and partial occlusion of target objects in controlled environments, but real-world object recognition remains an open problem. Object recognition methods typically employ a training phase, in which a catalog or database of instances of specific objects to be recognized is created. In the testing phase, one aims towards identifying new instances of these objects. Object category recognition refers to the problem of possibly recognizing previously unseen objects, given that some objects belonging to the same category were available during training. For example, to build an object category detector to detect cars in a street scene, we usually train it with a number of images of different types of cars. The training aims at learning features extracted from different types of cars such that the trained system generalizes well enough to identify previously unseen models and makes of cars in the test images.

Traditional methods of object category recognition rely on features extracted from the color image obtained by a typical RGB camera. The advent of low-cost depth cameras, like the Microsoft Kinect has opened up a number of possibilities to use depth information to extract more informative features. Hence, one of the major research questions when dealing with both color and depth (RGB-D) information is to investigate which features best capture the knowledge about the target object category. The first contribution of this paper is to identify a set of features that are

[*] Corresponding author. Tel.: +49 8153 281648.
*E-mail addresses:* haider.ali@dlr.de (H. Ali), faisal.shafait@uwa.edu.au (F. Shafait), eirgiann@csd.auth.gr (E. Giannakidou), avakali@csd.auth.gr (A. Vakali), nadia.figueroa@nyu.edu (N. Figueroa), t.varvadoukas@nyu.edu (T. Varvadoukas), nmav@alum.mit.edu (N. Mavridis).

0921-8890/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.robot.2013.10.001

able to robustly capture shape as well as color information for object category detection. When capturing images in a natural environment, each scene image might contain multiple objects. The aim of scene labeling is to automatically identify which object categories are present or visible in the scene and locate them accordingly.

In natural environments, different object categories usually co-occur (like car and pedestrian). Hence, when recognizing a particular object, not only the features extracted from that object are useful, but also identities of other objects visible in the scene might be informative. For example, one might often find a keyboard and a mouse next to a computer monitor; thus, these objects are often co-occurring, thereby jointly constituting an object-to-object context. One way of incorporating information about possibly noisy identities of other objects in the scene is to jointly optimize the identities of all detected objects using object classifier's ranked or probabilistic beliefs about them. The second contribution of this work is the use of a Markov Random Field (MRF) model to take into consideration natural or expected co-occurrences of different objects to refine the results of object category classifier. Moreover, we investigate social media mining as a rich source for extracting natural co-occurrences of objects and analyze its transferability when used to aid scene labeling in artificially created benchmark datasets.

Furthermore, apart from combined depth-color and object-to-object context, one more possibility that has come up recently is to enhance training sets for recognition, by using depth-and-RGB images or 3D model object databases available online on the internet (see Appendix A). However, as we shall see, it is not always the case that such online information can enhance recognition, as this highly depends on whether such training sets generalize well to the specific testing set that we are aiming towards. Thus, in order to tackle the problem of context-aware object category recognition we intersect and exploit two sources of information: not only a multi-view RGB-D object dataset but also social media metadata; and we use them to feed an object category classifier. This is achieved by processing each source of information independently and then combining it with a probabilistic graphical model (Fig. 1).

We construct a multi-class feature-based linear classifier. This classifier is trained with combined visual-shape features from multi-view instances of an RGB-D object dataset. The probabilistic output of this classifier is then combined with object co-occurrence probabilities derived from the statistics of relationships between object category labels. These two dependencies are jointly modeled with a MRF to provide context-aware object beliefs. Thus, this paper presents several key contributions for contextual object category recognition. These contributions consist of:

1. We use a classifier based on a novel fusion of feature vectors (the *VFH-Texton*), one of which is appearance based, and one is 3D shape based.
2. We furthermore derive a novel object-to-object context MRF model, based on crowd-sourced Flickr label co-occurrence data.
3. We investigate the convergence of the model's parameters as a function of Flickr's sample size.
4. Finally, we train the system on the University of Washington RGB-D Object Dataset [1] but then we do not test only on it, as all other existing work do. Instead, we assess the performance of our system in the real world: and thus we provide results on the generalization ability in a much harder real-world dataset, namely the NYU Indoor Scene Dataset [2], which has seldom been explored in the literature. Furthermore, we illustrate enhanced performance through the use of the MRF, which is highly correlated to the quality of object co-occurrences.

All of the contributions listed above clearly illustrate the power of our proposed method towards making real-world household object recognition using economical sensors a reality, which is an ability that would be highly beneficial for a number of applications, ranging from robotics to assistive devices for the blind and beyond. This paper is organized as follows. Section 2 provides background for each component of our contextual object category recognition system. A description of the extracted visual-shape feature histogram and details of the classifier are presented in Section 3. In Section 4, a discussion on learning context from social media is presented. Details of the MRF context model that uses the object classifier probabilities are given in Section 5. In Section 6, the evaluation of our system on both RGB-D Scene Dataset and NYU indoor scene dataset v2 is described. Finally, we present the discussion and conclusion of our work in Sections 7 and 8.

## 2. Background

In this section, we present the state-of-the-art of each component of our contextual object category recognition system.

### 2.1. Object category recognition in RGB-D images

In this subsection, we present a review of the state-of-the-art of combined visual appearance and shape feature descriptors extracted from RGB-D images and their use in object category recognition. In the past years, RGB images have been used to extract visual appearance information from objects and scenes with the well-known SIFT (Scale Invariant Feature Transform) [3] and SURF (Speeded Up Robust Features) [4] algorithms. However, several problems are encountered when applying them on real-world environments, such as variance in illumination, textureless objects, occlusion and surface reflectance, which cause these descriptors to perform poorly. With the release of low-cost RGB-D sensing devices, such as the Microsoft Kinect and the RGB-D Camera developed by Primesense, the robotics and computer vision communities have focused on developing techniques that exploit both color and depth information from scenes to increase the capabilities of object recognition and 3D modeling tasks [1,2,5]. This is specifically useful for object category recognition in scene labeling—the representation of an object can be described in terms of both visual appearance (from RGB image) and shape (from depth image) information. Recent work covering the challenges of object labeling in indoor scenes using RGB-D sensors has been reported by Ren et al. [6]. The classical object category recognition approach in the RGB domain is to train classifiers with extracted features from object models. The main challenge in using RGB-D data for object category recognition is the combination of the visual appearance and shape information in a fully descriptive and discriminative feature descriptor.

A comparative evaluation of 3D features implemented in the open source Point Cloud Library (PCL) [7] for object and category recognition has been presented by Alexandre [8]. One of his main conclusions was that adding color information to the feature vector increased recognition performance, naming CSHOT (Color-SHOT) [9] and PFHRGB [7] as the features with the best recognition accuracy over all. Figueroa et al. [10] performed a similar evaluation on a smaller set of features from PCL for a 3D registration application. It was concluded that by exploiting color information, such as using the CSHOT feature vector, high accuracy 3D registration results were obtained. More recent research efforts from Blum et al. [11], Tombari et al. [9], Ren et al. [6] and Lai et al. [12,13] have also shown that the combination of visual appearance and shape features result in a more discriminative description of the object or scene, compared to solely using shape or appearance. In Table 1, we present a comparison of existing methods *(in the literature)* that

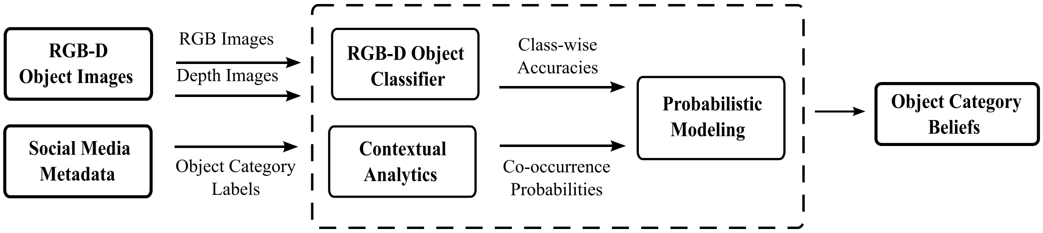


**Fig. 1.** Proposed system architecture for context-aware object category recognition.

**Table 1**
Literature overview of combined visual appearance-shape feature descriptors.

| Feature descriptor | Data representation | Feature extraction method |
|---|---|---|
| Convolution *k*-means [15] | RGB-D image | Histograms of triangular responses on the image channels |
| EMK-Spin + EMK-SIFT [1] | RGB + Depth | Concatenate EMK features from RGB and depth images |
| BRAND [16] | RGB + Depth | Encode intensity + shape change in a binary bit string |
| CSHOT [9] | Textured 3D point clouds | Concatenate shape + color histograms from 3D points using L-1 norm between normals and color triplets |
| PFHRGB [17] | Textured 3D point clouds | Concatenate shape + color histograms from 3D points using a triplet of angles between normals and colors |
| MeshHOG [18] | Textured triangular mesh | Vertex-based 3D histogram with color information |
| VOSCH [19] | Voxelized textured point clouds | Additive histogram from voxelized shape and color features |

extract combined feature descriptors from RGB-D data. The differences between these descriptors are two-fold: (i) the representation of the RGB-D image (ii) the feature extraction method used on this data representation. We provide more details in [14].

Based on the overview in Table 1, we have proposed a novel combination which exploits the power of two highly discriminative global features, i.e. the Viewpoint Feature Histogram [20] and the texton histogram [21]. The former is a global view-point based feature histogram extracted from the 3D point cloud representation of an object and the latter uses oriented Gaussian filter responses to retrieve texture information. We feed these features to a multi-class SVM classifier for object category recognition. Furthermore, we augment the classical object category recognition approach by combining the output of the RGB-D object classifier with a probabilistic representation of spatial context between objects in a scene (Fig. 1).

### 2.2. Context mining from social media analytics

We use social-media metadata to learn object contextual relationships from an ever growing dataset of tagged images from every-day scenes. Social media have attracted growing research interest since their emergence [22,23]. Indeed, the great amount of content along with the readily available metadata constitute a rich information source for analyzing and extracting useful knowledge, such as tag and user relations, rich related content, trends and dynamics in the current world. Specifically, a number of efforts utilize contextual information, in Flickr, for photo categorization regarding a specific object [24], location [25,26] or event [27,28]. These efforts couple object recognition methods with social-media metadata analysis (the latter is used to improve performance of the former). In all of the aforementioned approaches, it is shown that the use of social-media metadata as contextual knowledge significantly improves the analysis task at hand. It is worth mentioning that clustering is applied in most of the aforementioned approaches as a preliminary step for grouping related content. In the social-media literature, in general, clustering is a typical procedure used to tackle the inherent limitations of user generated content, e.g. ambiguous terms or synonyms [29].

There has been a considerable amount of interest in the research community in analyzing dynamical aspects of social media for discovering tag usage regularities and patterns. Remarkably, it seems that, despite the arbitrary tag usage, the tag space in a social-media platform is not a chaotic landscape; instead stable tag patterns emerge upon usage. More specifically, Golder and Huberman have shown that social tagging systems possess the dynamics of complex systems, exhibiting stable patterns in a resource's tag proportions over time [22]. They attributed this behavior to the dynamics of a stochastic urn model proposed in [30]. Specifically, they simulate the tagging task as a random colored ball selection from an urn and resemble tag reuse to an insertion in the urn of an additional ball of the same color. Furthermore, the same authors discovered regularities in tag frequencies, kinds, and manners in which tags are used, as well as burst of popularity of certain tags. In addition, in [31] it has been shown that tag distributions describing different resources converge over time to stable power law distributions. Once such stable distributions emerge, one can also examine the correlations between different tags. Both aforementioned approaches demonstrate their findings empirically on data collected from del.icio.us.

### 2.3. Contextual modeling for object recognition

The idea of using spatial context relies on the fact that certain objects typically occur in specific environments or are likely to be close to other specific objects. Therefore, it becomes a necessity to model these *object–object* relationships. Recent research efforts in computer vision (e.g. [32–34]) focus on how contextual relationships may improve object detection and recognition—especially in indoor environments which is relevant to the topic discussed in this paper. A classic representation of such spatial knowledge is through categorical clauses. Alternatively, local features statistics are exploited to generate contextual cues in vision problems, in order to identify real-world scenes (global context) and then focus on specific scene regions where the object is most likely to be found [35,36]. The notion of context can also be seen as dependencies between random variables, hence there is much work in modeling such dependencies with probabilistic graphical models [37,38]. Our work is mostly relevant to [38], where a Markov Random Field is used to model the co-occurrence relationships of objects and the object classifier's class-wise accuracies.

Other works that also couple object recognition with social-media metadata usage are [39,40]. In [39] the authors use word co-occurrence statistics from contextual data in Flickr to estimate *textual query—visual detector* similarity, in an effort to tackle the semantic gap between the detector's low-level features and the query's high-level semantics. In this work, we provide a detailed analysis on why calculating word relatedness by using web derived information is better than employing static semantic structures (e.g. ontologies, dictionaries). In [40], social tags and content from Flickr are used to train classifiers for object recognition. The results are promising and demonstrate a relatively easy and less time consuming way to train classifiers, compared with the manual training.

Based on the promising results of the social-media contextual information usage presented in Section 2.2, we propose using the crowd-sourced derived co-occurrence relationships in Flickr as parameters in a MRF model, to jointly utilize the RGB-D object classifiers' output along with object context for an object recognition task. To demonstrate the stability of the model, a convergence analysis of the tag co-occurrence in Flickr is presented in Section 4. To the best of our knowledge, the dynamics of tag co-occurrence as a function of the dataset size have not been explored yet in the literature, which is a crucial factor for exploiting real-world data and transferring it to synthetic data.

## 3. RGB-D object category recognition

For object category recognition we use a supervised learning approach, where feature vectors are extracted from annotated images of the UoW RGB-D Object Dataset and given to a linear classifier with their respective ground truth object class labels. The classifier then learns a model to categorize different object classes, and is tested on feature vectors extracted from segmented images of RGB-D indoor scenes (from two datasets: UoW and NYU).

Next, we present our proposed feature vector extraction approach based on combined visual and shape descriptors and describe the linear classification problem formulation. Furthermore, we describe the training and testing datasets and the pre-processing steps needed to use them in our experiments.

### 3.1. RGB-D feature extraction

In order to facilitate classification whilst considering both shape and appearance of an object, we propose a novel combination of global appearance and shape-based feature vectors, we call it the *VFH-Texton* (Fig. 2).

Our feature vector is based on a highly discriminative and efficient 308-D feature descriptor extracted from the 3D point cloud representation of the object, the Viewpoint Feature Histogram (VFH) [20]. The VFH encodes geometry and viewpoint and is robust to surface noise and discontinuities generated from stereo or Kinect data. Furthermore, it has been shown to reach higher recognition rates than the well-known Spin Images [20]. This feature descriptor is constructed by binning geometric relations between each point $p$ in a point cloud $P$ and the point cloud's centroid $c_P$. Additionally, the angle between the translated central viewpoint direction and the surface normal of each point is also binned, capturing the variance in viewpoint directions. We combine the VFH with a 539-D texton histogram feature vector which uses oriented Gaussian filter responses to retrieve texture information [41]. This feature vector is composed of 14 different features that capture relevant information on shape, color, texture and position of an object in the RGB image space [21]. It has been shown to perform extremely well in multi-class object recognition tasks. Thus, by combining (through concatenation) these two global feature vectors (VFH and texton), we generate a 847-D feature vector which represents both visual appearance and shape of the object. For numerical consistency, the VFH histogram is scaled over the total number of points resulting in a combined histogram with a range of [0,1].

The extracted 847-dimensional VFH-Texton feature vector is used to represent instances of segmented objects for training and testing of multiple object categories from the selected datasets. In Section 6.1, our proposed feature vector is evaluated against the EMK-SIFT + EMK-Spin feature vector proposed by Lai et al. [1]. The experimental results show that the use of our VFH-Texton increases performance for object classification.

### 3.2. Problem formulation—linear object classification

Considering our proposed feature vector and our training dataset (see next section), each object class instance has 847 features (VFH-Texton) and each object class has roughly 2000–5000 instances (Fig. 3). The number of instances per object category depends directly on the number of views provided by the UoW RGB-D object dataset. Our feature instances are extracted from multiple views of an object spun around on a turntable with the camera at different heights—this yields roughly 500 views (i.e. feature instances). However, each object category contains several types[1] of objects. Thus, if the *coffee mug* category has 8 different types of coffee mugs it will contain roughly 4000 views, whereas the *cereal box* category, which has 5 different types, contains roughly 2500.

Within our framework, we need a classifier that provides a probabilistic output for classification; i.e. not just labels of the identified classes but also the probabilistic belief of that label. In the binary classification scenario, we chose a linear maximum margin classifier based on Logistic Regression (LR), whose loss function is derived from a probabilistic model [42]. Given a set of instance-label pairs $(\mathbf{x_i}, l_i)$ where $\mathbf{x_i} \in \mathbb{R}^{847}$ is the RGB-D feature vector of the $i$th object, each instance is assigned to a binary relevant label $l_i \in \{1, -1\}$ for a each of the $K$ given number of object classes $O = (o_1, \dots, o_K)$ through our model. For multiple object classification in a scene, we use a *one-against-one* linear-kernel multi-class method [43]. We provide more details in [14].

### 3.3. Training and testing datasets

The dataset used to train our probabilistic classifier is the RGB-D Object Dataset collected by the University of Washington (UoW) [1]. This dataset contains color and depth images for 300 objects organized into 51 categories taken from multiple views. The objects at hand are common household and office related objects such as fruits, office items, kitchen accessories, etc.

The system is tested on two different datasets. The first one is the RGB-D Scene Dataset which contains annotated scenes containing objects from the RGB-D Object Dataset by the UoW (Section 6). This dataset does not provide per-pixel segmentation of the annotated objects; only the bounding box of each annotated object is provided. Therefore, a pre-processing step on the scenes had to be applied to segment the objects in the color and depth images (Section 3.3.1). The second dataset is the NYU indoor scene dataset V2 [2]. This dataset provides per-pixel accurate object segmentation masks for all of the scenes, therefore no segmentation is needed. However, it is more challenging than the RGB-D UoW dataset due to the difference in viewpoint, appearance, and proximity of the objects (Section 6.4).

#### 3.3.1. RGB-D scene dataset object segmentation

As mentioned previously, the RGB-D scene Dataset does not provide pixel-accurate object segmentation masks. Therefore, to test our classifiers on the scene data, we generated ground truth segmentation masks by combining two sources of information: (i) ground truth bounding box annotations and (ii) output of 3D segmentation on the scenes. For example, we use the annotated RGB image and the 3D representation of the scene as shown in Fig. 4(a)–(b).

[1] In the UoW RGB-D Object dataset *instances* is used for the different object types per category [1]. Here, to avoid confusion we use *types* to represent the different objects.

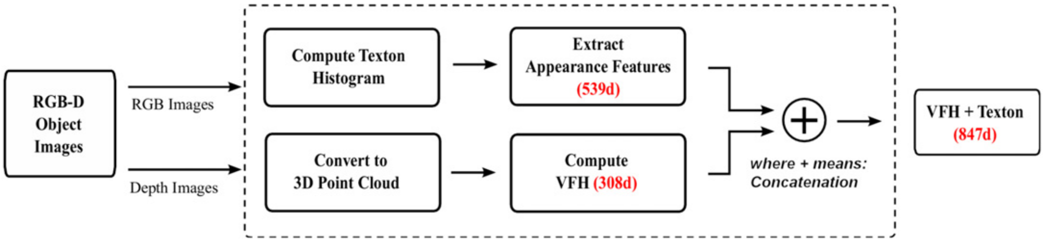
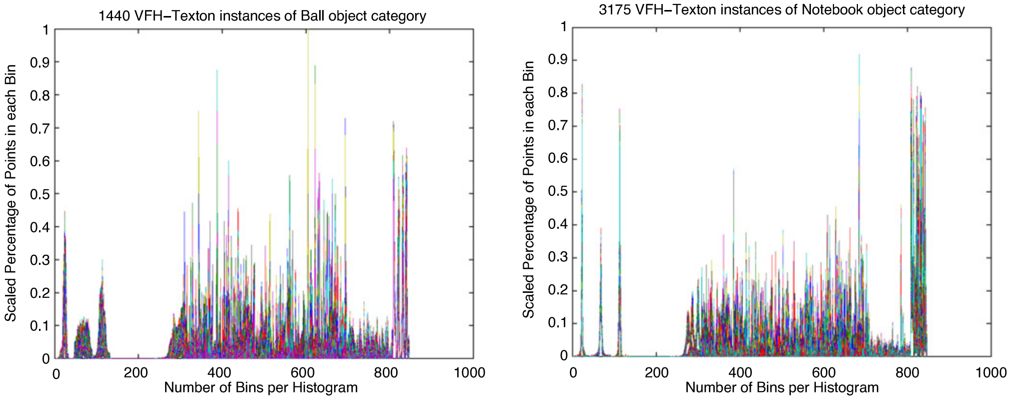


**Fig. 2.** Feature extraction pipeline: two types of feature histograms (VFH + Texton) are extracted from both RGB and depth images to generate a 847-D feature vector.



(a) *Ball* VFH-Textons.

(b) *Notebook* VFH-Textons.

**Fig. 3.** Superposition of VFH-Textons for two object categories: (a) Ball with 1440 object instances and (b) Notebook with 3175 object instances. Different colors represent different VFH-Textons. In the horizontal axis of the diagrams, components 1 to 307 correspond to the VFH components, while 308 to 847 correspond to the Texton components.

The annotated images from the scene do not include all of the existing objects, and furthermore the bounding boxes cut off the objects in some cases (i.e. flashlight and cup). Also, the 3D point cloud has its limitations—mainly the discontinuities on the surfaces of the objects lead to over-segmentation for some instances. Nevertheless, the following algorithm attempts to exploit the benefits of each source as much as possible whilst alleviating the above issues. The segmentation algorithm consists of the following steps.

1. *Planar model fitting*: The first step for segmenting the objects in a scene (*with the assumption that they lie on a table*) is to find the table-top area where these objects are located. We use a Random Sample Consensus (RANSAC [44])-based method to iteratively estimate from a set of 3D points of the scene parameters of the mathematical model of a plane, $ax + by + cz + d = 0$, where $a, b, c$ are the normalized coefficients of the $x, y, z$ coordinates of the plane's normal and $d$ is the Hessian component of the plane's equation. The largest fitted plane is segmented from the point cloud and represents the object-supporting surface (i.e. table or counter) of the scene (Fig. 4(c)).

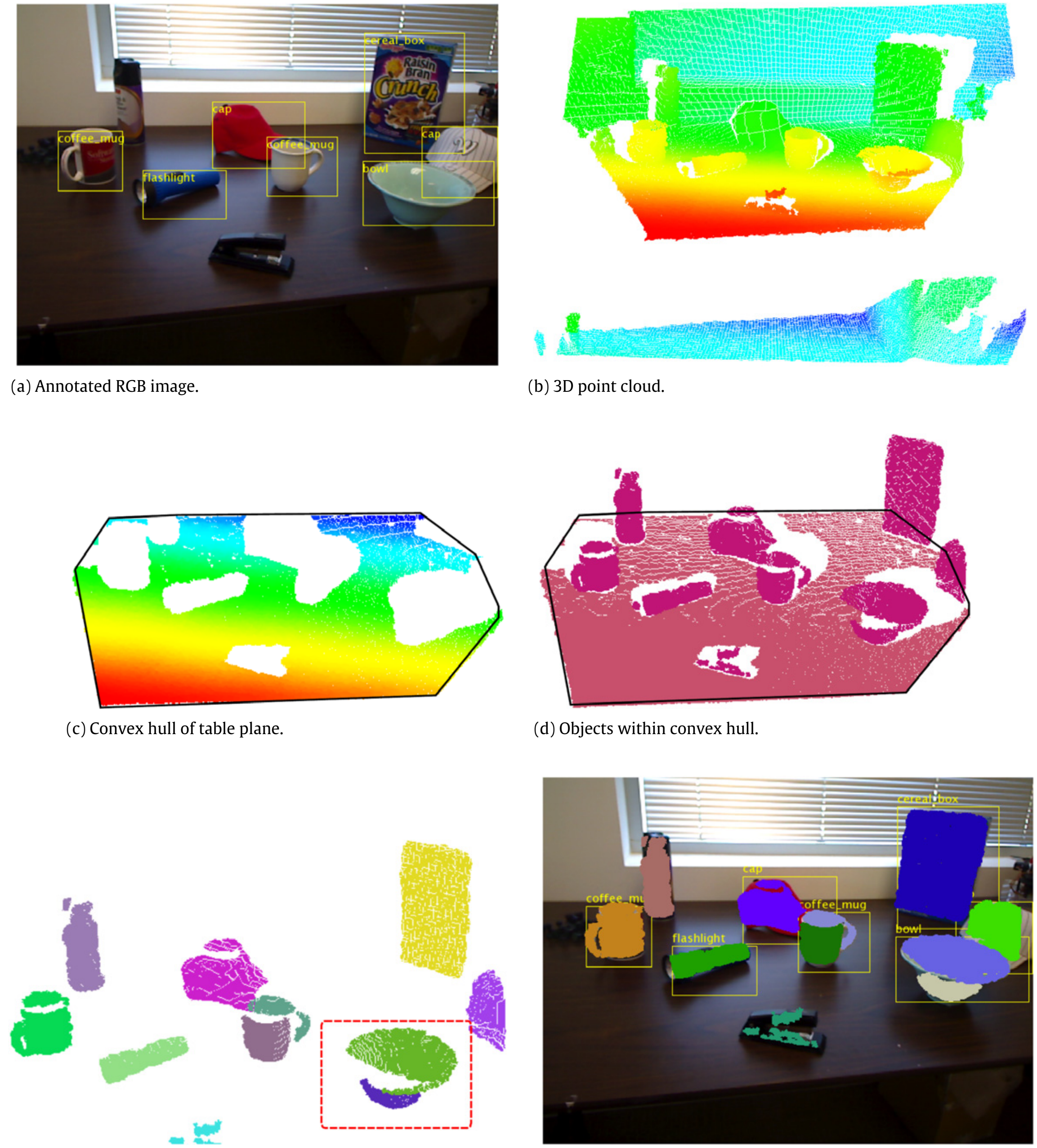
2. *Object extraction within convex hull of a plane*: At this stage, the plane of the table-top has been identified, however our interest is extracting the set of points that lie on top of this plane and are within the convex hull of the plane (Fig. 4(d)).

A 3D convex hull polygon is computed for the set of points formed by the plane using the Quick Hull algorithm introduced in 1977 by W. Eddy and in 1978 by A. Bykat. As the table plane is inclined with respect to the origin of the camera, we must find the local reference frame of the table in order to locate the 3D points that are "on top" of the table (i.e. the objects). To simplify the object extraction procedure, we constrain the local reference frame of the table to have its $z$-axis parallel to the surface normal $n_i$ directions of the plane and the plane must be orthogonal to the $z$-axis and parallel to the $x$–$y$ plane of the local coordinate system. The surface normals $n_i$ of every point $p_i \in P_{\text{plane}}$ of the plane have the same orientation throughout the whole surface. To achieve the transformed plane, we need to find the transformation $T$ that will transform the point cloud from camera world coordinates to local plane coordinates. Once the point cloud is in the local reference plane, the objects are easily extracted by offsetting the 2D convex hull in the $z$-direction of the local reference plane. In order to construct this transformation matrix $T$, we find a unique 3D rotation $R$ that will rotate the $z_{\text{direction}}$ of the plane (i.e. the normal direction) into $(0, 0, 1)$ ($z$-axis) and the $y_{\text{direction}}$ of the plane orthogonal to the $z$-axis. More details can be found in [14]. The procedure to extract the objects from the full point cloud of the scene is listed in Algorithm 1.

**Algorithm 1** Object extraction procedure

```
Input: T (rigid transformation), $P_{\text{plane}}$(point cloud of the segmented plane), $P_{\text{scene}}$(full point cloud of the scene)
Output: $P_{\text{objects}}$ (point cloud representing the objects on top of the table)
    $P^*_{\text{plane}} = T \cdot P_{\text{plane}}$
    $P^*_{\text{scene}} = T \cdot P_{\text{scene}}$
    $3DConvexHull \leftarrow constructConvexHull(P^*_{\text{plane}})$
    $Extruded3DHull \leftarrow extrude3DConvexHull(3DConvexHull)$
    $P^*_{\text{objects}} \leftarrow extractPointswithinHull(P^*_{\text{scene}}, Extruded3DHull)$
    $P_{\text{objects}} = T^{-1} \cdot P^*_{\text{objects}}$
```

Initially, $P_{\text{plane}}$ and $P_{\text{scene}}$ are transformed by $T$ so that the plane is orthogonal to the $z$-axis. Then, the 3D convex hull polygon of $P_{\text{plane}}$ is extracted and extruded in the $z$-direction to create a 3D bounding shape around the 3D plane of the table top. The points within this 3D bounding shape are extracted and transformed back to the original world coordinate system, resulting in a point cloud containing only the objects on the table top (Fig. 4(e)). We transform the extracted objects back to the world coordinate frame since they have to be projected back to the image plane in step 4 below.



(a) Annotated RGB image.

(b) 3D point cloud.

(c) Convex hull of table plane.

(d) Objects within convex hull.

(e) Clustered objects.

(f) Projected clusters.

**Fig. 4.** Six stages of segmentation from an example test scene of the RGB-D dataset: (a) Annotated RGB image, (b) 3D point cloud of the scene, (c) Largest segmented plane resulting from planar model fitting, (d) Objects lying within convex hull, (e) Clustered objects (each color represents a different cluster), an instance of over-segmentation is shown within the red bounding box, (f) Projected clusters with annotations.

3. *Euclidean clustering*: To segment the $n$-objects from the resulting point cloud, we extract $n$-clusters from $P_{\text{objects}}$, by applying Euclidean clustering in 3D coordinates $(x, y, z)$, which consists of extracting clusters of data that lie within a certain search radius $(r)$ restricted to a user-defined spatial cluster tolerance $(d)$, that is computed as a measure in the L2 Euclidean space. Thus, for every point $p_i \in P_{\text{objects}}$ a set of nearest neighbors $nn_k^i \in NN^i$ within a sphere of radius $r < d$ is searched for using a KD-tree representation of the $P_{\text{objects}}$ [45]. The neighbors $nn_k^i$ that lie within the spatial cluster tolerance are added to the cluster, if they have not been processed before (i.e. were not a neighbor for a previous query point $p_i$). This procedure is repeated for all points $p_i \in P_{\text{objects}}$ until all of them are part of a cluster. The implementation of this approach is found in the open source Point Cloud Library (PCL) [7]. We provide the algorithmic steps in [14]. The resulting segmentation is shown in Fig. 4(e). As can be seen, due to major discontinuities in the depth image for some objects (cup and bowl) an over-segmentation (Fig. 4(e)) is present, to be dealt with in step 4.

4. *Back-projection and annotation matching*: The extracted clusters from the point cloud of the scene have no labels, meaning we do not know what object they are—but we do know that they are objects. Therefore, to find the matching annotation the clustered 3D points are projected back to the RGB image and overlaid on the bounding box (Fig. 4(f)).

A binary mask is created for each projected cluster. The annotated bounding boxes are projected on each binary mask. The mask that contains a binary component of a minimum area near the centroid of the bounding box is chosen as the pixel-wise mask for that specific annotation.

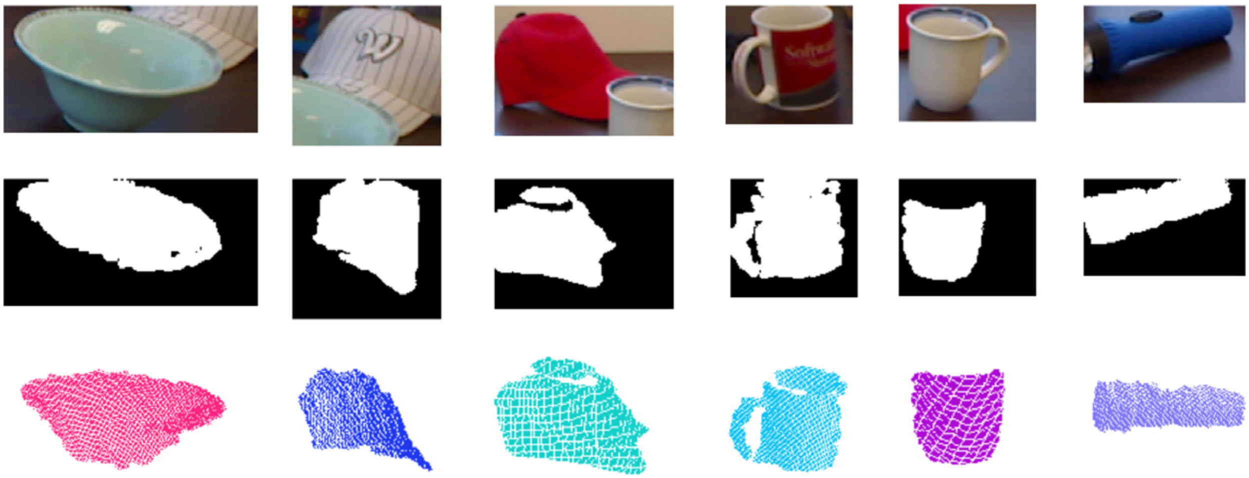


**Fig. 5.** Results of the segmentation algorithm: (top row) segmented RGB image, (middle row) pixel-accurate binary mask, (bottom row) segmented point cloud. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The proposed segmentation algorithm generates three data sources that describe the segmented object: (i) RGB image (ii) binary mask and (iii) point cloud. The RGB image is used in our feature extraction algorithm to compute the SIFT descriptors which are used for the EMK-SIFT feature. Both the RGB image and the binary mask are used to compute the texton histogram and the point cloud is used to compute the VFH descriptor. In Fig. 5, we show the results of our segmentation algorithm. As can be seen, for some objects the projection of the segmented point cloud from 3D world coordinates to the image plane is the accurate representation of the object. Nevertheless, it is possible to generate masks with a large amount of missing pixels—such as the stapler, which is not annotated. This happens due to missing pixels in the depth image generated from reflective materials or external noise. When objects contain this type of noisy measurements, it presents a challenge for object category recognition, because the generated point cloud of the segmented object has missing information (compared to the point clouds of the soda can model from the RGB-D object Dataset). Furthermore, when over-segmentation occurs, (such as the coffee mug in the center of the scene or the bowl) our algorithm considers only the largest segment that is localized in the center of the bounding box. Morphological operations on the binary mask could be applied to connect these separated clusters that describe the same object, however if this is done the risk of connecting distinct objects is increased—due to the proximity between them (see coffee mug and cup or bowl and cup). An idea to overcome this over-segmentation is to pre-process the depth image in order to smooth out the surfaces and obtain fully connected segmented objects prior to the back-projection step—this is being investigated in our current research. Even though the segmentation is not perfect, it is useful and since our object classifiers rely not only on shape features but also on visual features—they are robust to these inconsistencies of segmented objects, which represent the challenges in real-world applications.

## 4. Learning context from social media

The ease of sharing and annotating multimedia content through social-media sites feeds the Web with enormous scales of image and video data along with descriptions of the objects located within. Indeed, sites like Flickr accommodate image corpora which are populated with hundreds of user tagged images on a daily basis Apart from the large amount of content, an important information source is the contextual data provided by users to describe and/or annotate such content. Contextual data may be semantic, spatial and/or temporal descriptors. These additional data associate images and videos in such sites with specific semantic concepts, places and/or time periods. Typically, semantic descriptors in social-media sites are given in the form of tags, i.e. freely chosen textual descriptions. Research on tag usage has shown that tags overwhelmingly identify the topics of their referents [22]. In addition, by analyzing semantic descriptions, which is the focus of this section, implicit knowledge can be inferred about how often some object categories co-occur in photos or in videos. Based on the assumption that if 2 objects co-occur in a photo, they are located in the same real-world scene, in this section, we address the problem of extracting object categories' contextual relationships in real-world scenes by analyzing tag co-occurrence in images from social media. For example, in terms of common everyday objects, "table" is expected to co-occur more frequently with "chair" rather than "toothbrush". Learning co-occurrence relations among object categories is expected to improve the recognition task at hand, since they can be used as prior knowledge and bias the classifier, after some first detections, towards more probable object categories, over others not so likely.

In this section, we aim at assessing object categories' contextual relationships by analyzing tag co-occurrence in social-media sites, rather than using a textual corpus (e.g. dictionaries, documents). Tag co-occurrences capture users' common perception for objects when interacting in social-media platforms. Furthermore, tags constitute a more dynamic indication of word inter-connection, when compared with the static relationships assigned in a dictionary, like Wordnet. It should be noted that semantic relatedness alone is not always an indicator of objects actual co-occurrence in the real environment. For example, "wheat" and "cereal" are semantically related but quite unlikely to co-occur in a same scene, whereas "cereal" and "bowl" are not semantically close but they are quite likely to be found together. Assessing contextual relationships via analysis of social content manages to capture such relationships, as will be shown below. Furthermore, since social media (like Flickr) accommodate images depicting a huge variety of object categories, the statistical analysis of contextual relationship among any of these categories can be inferred. Therefore, in essence, there is no limitation on the number of object categories used.

Furthermore, to account for the great variance in the users' tagging habits, we investigate the convergence of the co-occurrence parameters as a function of the Flickr dataset size used. To this end, we present a thorough experimental evaluation, where we check convergence by defining random variables that capture the context parameter distribution and iteratively calculating and checking the mean, $\hat{\mu}$, and variance, $\hat{\sigma^2}$, values of these variables for Flickr datasets of increasing size.

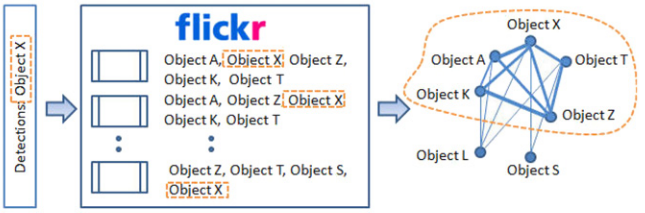


**Fig. 6.** Learning object contextual relationships from Flickr tags. In the object recognition scenario an already detected object serves as input (Object X). The proposed SMCS measure is used to select objects that are more likely to co-exist with Object X.

### *4.1. Problem formulation—learning co-occurrence relationships*

Let $O_i$, $i = 1, \ldots, N$ be $N$ object categories, and $o_i$, $i = 1, \ldots, N$ be their respective tags (i.e. labels). Given $N$ datasets $\mathcal{D}_i(s)$, each of them containing $s$ social-media photos tagged with $o_i$, and $\mathbb{D}_s \equiv \bigcup_{i=1}^{N} \mathcal{D}_i(s)$, we address the problem of identifying the object category $O_y$ that is most likely to co-occur with another object category $O_x$ in an indoor environment through tag analysis in $\mathbb{D}_s$. Intuitively, based on the large number of photos in social media sites, we claim that the possibility of observing object categories $O_x$ and $O_y$ together in a real-world scene is highly reflected on the degree of the pairwise appearances of their corresponding tags $o_x$ and $o_y$ in photo annotations. Hence, we cast our problem as a tag similarity problem based on co-occurrence in a social-media derived dataset $\mathbb{D}_s$. We define *co-occurrence similarity*, *CS* between 2 tags $o_x$ and $o_y$ in a $\mathbb{D}_s$ as

$$CS_s(o_x, o_y) \equiv \frac{|o_x \cap o_y|}{|o_x \cup o_y|}, \tag{1}$$

where the numerator expresses the number of images in $\mathbb{D}_s$, in which tags $o_x$ and $o_y$ co-occur, whereas the denominator equals to the number of images that were tagged with either $o_x$ or $o_y$.

Then, we apply an exponential kernel on *CS* to obtain the following measure named *Social Media Contextual Similarity*, that captures the co-occurrence between 2 object categories $O_x$, $O_y$ in an indoor environment.

**Definition 1** (*Social Media Contextual Similarity*). Suppose we have two object categories $O_x$, $O_y$ and the corresponding tags $o_x$, $o_y$. The Social Media Contextual Similarity **SMCS** between $O_x$ and $O_y$ is defined as

$$\mathrm{SMCS}(O_x, O_y) \equiv e^{-\frac{CS(o_x, o_y)}{\tau}},$$

where the selection of $\tau$ value is made empirically as the average pairwise *CS* of a randomly pooled set of tags. This empirical method has shown promise in kernel-based machine learning tasks [46].

Fig. 6 shows the general framework of extracting contextual relationships from Flickr tags in the object recognition scenario. More specifically, we query Flickr (via its API) with the object category label $o_x$, to get $s$ photo annotations that contain $o_x$ as tag, and construct $\mathcal{D}_s(x)$. We construct, the same way $\mathcal{D}_s(y)$ for any other object category label $o_y$ that can be detected by the classifier. Based on tag co-occurrence in these Flickr datasets, we can build a weighted tag graph, in which some tags are tightly related (high co-occurrence values), whereas others are loosely or not connected at all (small co-occurrence values). Since Flickr tags describe, amongst others, scene settings, we expect that tag co-occurrence captures the object correlation in a real-world indoor environment. Hence, we use the same graph with object categories as nodes and traverse it, to find object categories that are most likely to co-occur with a detected object category $O_x$.

However, intrinsic limitations of user generated content, such as lack of structure, tag ambiguity and use of synonyms, raise concerns about the quality of knowledge we can extract from social-media platforms, like Flickr. For example, in the case where ambiguous tags exist, the dataset will contain photos and co-occurrence relationships not related to the intended tag sense, "polluting", thus, our measure of co-occurrence. Furthermore, to reliably use SMCS as a model parameter, we need to check its stability, i.e. if it converges as the social-media dataset size $s$ changes. To tackle these issues, we investigate (i) tag co-occurrence statistical properties as a function of the Flickr *dataset size s*, and (ii) *purity*, where *purity*, pertains to the extent that the photos in a $\mathcal{D}_i(s)$ actually refer to the actual object category $O_i$. The latter is related to the ambiguity level of the label $o_i$.

### *4.2. Social media contextual similarity (SMCS) convergence analysis*

In what follows, we describe the datasets used in our experimentation and, then, address SMCS convergence as a function of dataset size and purity: two fundamental facets that are crucial to our analysis.

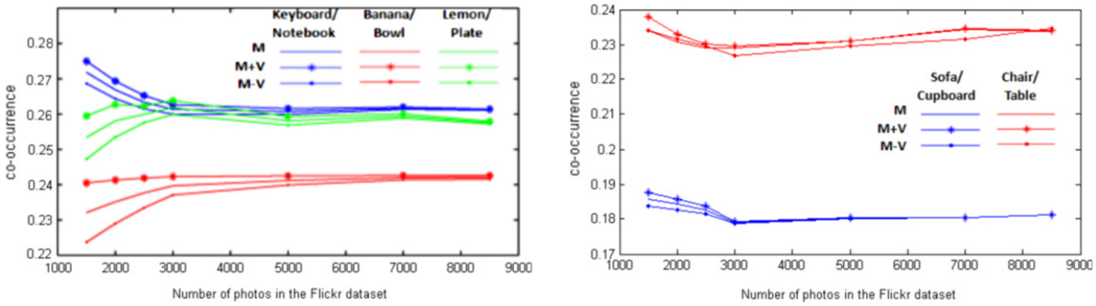
#### *4.2.1. Datasets*

Since our object recognition scenario's training and testing were built upon NYU [1] and RGB-D [2] datasets, we analyzed co-occurrence in a Flickr-derived dataset based on the 21 terms found in common in NYU and RGB-D datasets. Furthermore, to increase the soundness of the day-to-day object categories co-occurrence results, we used another Flickr dataset formed by the 75 terms found on the Berkeley dataset which contains common everyday objects and scenes [47]. Each dataset was created by querying Flickr an object category tag (term) via its API, to get the top 15 000 image annotations that contain this tag. This resulted in 587 967 images (Berkeley dataset) and 233 624 images (RGB-D + NYU dataset).

#### *4.2.2. Co-occurrence convergence as a function of Flickr's sample size*

Given a pair of object categories $O_x$, $O_y$, we aim at testing the hypothesis that SMCS($O_x$, $O_y$) converges as the number of photos in the dataset increases. The problem of estimating a parameter's convergence for larger yet unknown datasets is a difficult analytical problem. It amounts to developing a model to compute how fast a given parameter "learns" or improves its "fitting" to the data as a function of dataset size. In our case, this would mean to develop and learn $N^2$ models, where $N$ is the number of object categories. A natural way to study convergence as a function of dataset size is by building empirical scaling models called learning curves [48]. However this method applies to datasets with ground truth, where we can calculate the model error and check its convergence. In the problem of studying tag co-occurrence in social media there is no ground truth, hence no estimation for error can be realized. An empirical way to study convergence in this case would be by defining a random variable $X_{xy}$ on the tag co-occurrence event of $o_x$, $o_y$ and checking whether $X_{xy}$ converges to a specific value, as the dataset increases. It holds that if a variance of a random variable describing the next event (i.e. co-occurrence of $o_x$, $o_y$ in a dataset of increased size) converges to a small value, then a sequence of essentially random events can be expected to settle into a pattern [49]. Based on this observation, we aim at showing $X_{xy}$ convergence, by following an iterative process that can be broken down into two main tasks: (i) random subsampling of photos, to form a dataset $\mathcal{D}_x(s) \cup \mathcal{D}_y(s)$, and (ii) calculation of mean, $\hat{\mu} = E[X_{xy}]$ and variance $\hat{\sigma}^2 = E[(X_{xy} - \hat{\mu})^2]$. We repeat this procedure for a range of dataset sizes $s$ and test if the variance reduces with $s$ and, at the same time, $\hat{\mu}$ converges.

Let $X_{xy}$ be a random variable on the co-occurrence event of $o_x$, $o_y$ (as discussed earlier). Given a social-media derived dataset $\mathbb{D}_{T1}$,



(a) *RGB-D* + *NYU* dataset.

(b) *Berkeley* dataset.

**Fig. 7.** The convergence of $X_i j$ $\hat{\mu}$ and $\hat{\sigma^2}$ for 2 co-occurrence relationships. The vertical axis denotes SMCS$(i, j)$ values and the horizontal axis dataset size in units of photos added. (a) RGB-D + NYUs, (b) Berkeley.

where $T1 \gg s$, we construct by random subsampling 2 datasets of size $s$ from tags $o_x$ and $o_y$ and estimate their union, $\mathcal{D}_s(x) \cup \mathcal{D}_s(y)$. We repeat this process $k$ times. We calculate SMCS$(O_x, O_y)$ for each of the $k$ resulting datasets of size $s_1 < s_2 < \cdots < s_K$, respectively, where $s_K \ll T1$. Then, we compute the $\hat{\mu}$ and $\hat{\sigma^2}$, that is the mean and the variance values of $X_{xy}$ over the $K$ varying dataset sizes. To verify convergence, we depict how $\hat{\sigma^2}$ and $\hat{\mu}$ values vary over different dataset sizes.

Fig. 7(a) shows, indicatively, the relationships *Keyboard/ Notebook*, *Banana/Bowl*, and *Lemon/Plate* for various Flickr dataset sizes. Each curve represents a $X_{xy}$; as more tagged photos are added (horizontal axis), the mean of $X_{xy}$, $\hat{\mu}$ converges, while the variance of $X_{xy}$, $\hat{\sigma^2}$ (vertical axis) flattens out. Likewise, in Fig. 7(b) we observe the same patterns for contextual relationships *Chair/Table*, *Sofa/Cupboard* extracted from the Berkeley dataset: the co-occurrence values between two object categories tend to converge when the Flickr dataset reaches a size threshold.[2]

To determine whether the co-occurrence values between two object categories are indeed the same for the Flickr datasets above a specific size, we performed the Kruskal–Wallis test for the Keyboard/Notebook and Lemon/Plate co-occurrence results [50]. For the first null hypothesis that SMCS(*keyboard*, *notebook*) comes from the same distribution for the Flickr datasets $\mathcal{D}_s(i)'$ with size = 3000, 5000, 7000, we obtained chi-squared = 2.11, and $p$-value = 0.3474 (i.e. Prob > Chi-sq). For the second null hypothesis that SMCS(*lemon*, *plate*) comes from the same distribution for $\mathcal{D}_s(i)'$ with size = 3000, 5000, 7000, we obtained chi-squared = 8.44, and $p$-value = 0.1338. It is this chi-square statistic that was actually used to test the null hypothesis that SMCS$(O_x, O_y)$ converges as the number of photos in the dataset increases. The extremely high $p$-values that were obtained by the two implementations of the Kruskal–Wallis test are a strong indication that we cannot reject the null hypothesis. Hence, there is no statistical difference found in pairwise co-occurrence values drawn from Flickr datasets above a specific size.

### 4.2.3. Analysis of Flickr dataset purity effect

We refer to the *purity* of a social-media derived dataset $\mathcal{D}_j(s)$ as the extent to which the dataset contains photos related to the sense of the tag $o_j$ we would like to test. Due to the questionable tag validity and tag ambiguity in social-media, retrieval in these systems has, often, been a major issue. To tackle poor retrieval and obtain pure datasets, clustering has been proposed in the social-media literature, as an approach to overcome their intrinsic limitations, mentioned above, and group together related items around a certain sense [29,51]. Here, we employ a clustering approach that has been shown to be effective in (i) discovering different meanings in ambiguous tags, and (ii) grouping together tags and resources that refer to the same meaning [51]. In general, in the case of an ambiguous tag $o_j$, the dataset $\mathcal{D}_j(s)$ will contain a number of photos not related to the intended tag sense. In that case, the co-occurrence relationships will be extracted by considering all the senses of $o_j$, "polluting", thus, our measure of co-occurrence. For example, if we want to test the co-occurrence between "apple" and "orange", due to the fact that "apple" is an ambiguous tag and refers also to a computer brand (apart from the fruit), the photos tagged with "apple" and referring to computers will only increase the denominator of $CS_s(apple, orange)$. The decreased values of $CS$ for ambiguous tags will also cause a distortion on the SMCS values.

An empirical way to tackle this issue would be to define a boost parameter $\gamma$ that is applied on $CS_s(o_x, o_y)$ and return a re-enforced $CS'$:

$$CS'(o_x, o_y) \equiv \gamma * CS(o_x, o_y),$$

where the parameter $\gamma$ can be learned manually by observing in $\mathcal{D}_x(s)$ and $\mathcal{D}_y(s)$ of small size $s$ the percentage of annotations in which the ambiguous tags $o_x$, $o_y$ are being used in the sense we are interested in.

An approach we propose here, to tackle ambiguous tags, is to employ photo/tag co-clustering [51] in the Flickr datasets, to divide the photos referring to an ambiguous tag $o_x$ into semantic clusters based on each different meaning of $o_x$. To capture the latent topics that exist in a tag dataset we apply *factor analysis*. Factor analysis is used to find latent variables or factors among observed ones. Here, it is used to detect the latent meanings of ambiguous tags. Indicatively, Fig. 8(a), (b) show graphically factor analysis results for datasets of tags "keyboard" and "flashlight". An illustration for the analysis in $\mathcal{D}_{keyboard}(s)$ for the ambiguous tag "keyboard" follows.

Initially, we need to employ a selection process of the most distinguishing tags in $\mathcal{D}_{keyboard}(s)$, that will drive the clustering process, since, in practice, the number of tags in $\mathcal{D}_{keyboard}(s)$ may grow in large scale. We chose to use the very popular *Term Frequency–Inverse Document Frequency* (TF–IDF) statistic that reflects how important a tag is in $\mathcal{D}_{keyboard}(s)$ in relation to the dataset corpus—i.e. $\mathbb{D}_s = \cup\mathcal{D}_i(s)$, where $i$ is any object category—by estimating the frequency of a tag in $\mathcal{D}_{keyboard}(s)$ over the frequency of the tag in the $\mathbb{D}_s$. Thus, after performing a preprocessing, during which rare tags are removed, we select the tags with TF–IDF score above a threshold $\theta$[3] to be the observed variables in $\mathcal{D}_{keyboard}(s)$ and the attributes in the clustering process. Having selected the tag attributes, we proceed to the construction of a correlation matrix among the photo annotations in the $\mathcal{D}_{keyboard}(s)$ and the tag attributes. The similarity metric according to which the correlation matrix was built is based both on semantic similarity (drawn from WordNet) and tag co-occurrence, as described in [51]. This correlation matrix is the input to the factor analysis process,

[2] In our experiments this size was shown to be 3000 photos.

[3] $\theta$ is determined empirically. In our dataset, we used $\theta = 0.6$.

the results of which are shown in Fig. 8(a). As depicted in the aforementioned figure, the latent variables in this tag dataset is 2, based on the criterion of having eigenvalues greater than 0.4. It can be observed that there are two main groups of variables, each of which is concentrated in a different factor. Running the co-clustering algorithm [51] with number of clusters to be equal to the number of extracted factors, we see that the algorithm manages to identify groups of tags that appear to be near in terms of their semantic similarity. Indeed, in Fig. 8(a), the tags that have been assigned in the same cluster have been assigned the same color. Actually, the two factors could be interpreted as the two senses of the "keyboard" tag in the dataset, that is music-related and computer-related. Tags "piano" and "keyboard" are not colored, since they were wrongly assigned in the computer-related cluster. Finally, we set the $\mathcal{D}_{keyboard}(s)$ to be the photo annotations contained in the computer-related cluster. This way, we have a relatively pure Flickr dataset for our object recognition task.

In general, ideally, each semantic cluster corresponds to a tag sense. Then, the analysis is performed only on the annotations contained in the cluster that corresponds to the actual object category. The mapping between extracted clusters and object categories is based on cluster topics, as explained in [51]. Clustering does not generally achieve to separate perfectly the photos, so an error parameter *e* can be used in the SMCS metric, to represent the information loss, which is learnt upon training.

## 5. Contextual modeling

In this section, we describe how we use the crowd-sourced derived co-occurrence relationships (Section 4) as parameters to a MRF model, to jointly utilize the RGB-D object classifiers' output (Section 3) along with object context for an object recognition task. Context powerfully influences how humans recognize and locate object categories. Contextual information becomes extremely handy in cases of occluded or not easily seen objects, where a human visual perception system performs much better than computer vision systems do.

### 5.1. Problem formulation—contextual modeling using MRF

At this stage the goal is to combine the knowledge of object-to-object co-occurrence statistics with the output of a multi-class probabilistic classifier. In order to model these dependencies it is natural to use an undirected graph, since there is no clear directionality between the random variables. More specifically, with the aid of the graphical model we are interested in estimating the probability for a set of objects to co-occur in a scene given the beliefs of a classifier. Thus, we define a normalized distribution with the aid of an Markov network (MRF), by multiplying the local factors, and then normalizing it to obtain a valid probability distribution [52]:

$$P(o_1, o_2, \dots, o_n | b_1, b_2, \dots, b_n) = \frac{1}{Z} \prod_{i,j} \psi(o_i, o_j) \prod_i \phi(o_i, b_i) \qquad (2)$$

where $\psi$ denotes the co-occurrence statistics of the objects $o_i$ and $o_j$, $\phi$ the classifier's probabilistic belief for an object's detection and $Z$ is the partition function. The conditional joint probability in Eq. (2) is used to test the combination of the detected objects in a scene.

### 5.2. Computational efficiency

The choice of the MRF defined in Section 5.1, over a wide range of other probabilistic models, further offers computational efficiency due to the Markovian assumption. The assumption reduces the space of all the co-occurrence *n*-adic combinations to pairwise relationships.

In order to make the above model and its reduced number of parameters clearer, the following example is presented. Suppose that there are 3 objects $(o_1, o_2, o_3)$ in a scene and $k$ discrete classification categories $X_i, i \in [1, k]$. We want to calculate the probability that the objects are $(X_1, X_3, X_1)$ given the respective classifier's beliefs $(X_1, X_2, X_1)$:

$$\begin{aligned} &P(o_1 = X_1, o_2 = X_3, o_3 = X_1 | b_1 = X_1, b_2 = X_2, b_3 = X_1) \\ &\overset{\text{Eq. (2)}}{=} \frac{1}{Z} \prod_{i,j} \psi(o_i, o_j) \prod_i \phi(o_i, b_i) \\ &= \frac{1}{Z}\, \psi(o_1, o_2)\, \psi(o_1, o_3)\, \psi(o_2, o_3) \\ &\quad \times \phi(o_1, b_1)\, \phi(o_2, b_2)\, \phi(o_3, b_3). \end{aligned}$$

From the above analysis of the product it is obvious that the number of $\psi$-parameters increases if we model more than pair-wise relationships, whereas the $\phi$-parameters remain constant. To illustrate this, using the above example of 3 objects and $k$ categories $N_{\text{mark}} = 3k^2$ $\psi$-parameters need to be estimated, whereas without the Markovian assumption (that is to include the triad too) the parameters rise up to $N_{\text{nonmark}} = 3k^2k^3 = 3k^5$. For example, for $k = 4$, we have $N_{\text{mark}} = 248$, whereas $N_{\text{nonmark}} = 3072$. The difference is obvious even for a small number of categories and objects within a scene.

### 5.3. Objects' co-occurrence selection algorithm

The algorithm generates a set of different possible objects' combinations within a scene, and chooses the one with the maximum likelihood. It is natural to extract these combinations from the classifier's top beliefs. In order to restrict the exponential search space two parameters are introduced: (I) $k$: select the top $k$ beliefs for each object, and (II) *perc*: from these top $k$ select those whose belief is greater than *perc*. If very strict values are chosen for the threshold *perc*, then there is the possibility that no beliefs will ever pass the threshold. In that case, we complete the set so that it contains the top $k$ beliefs. The main reason is that the classifier is highly uncertain for this object, hence all of the possible decisions should be considered. This choice is meaningful when the percentage threshold is relatively low, since that would mean that the top decisions are probabilistically insignificant, and all of them should be taken into account. The final result is expected to be more dependent on the co-occurrence statistics. It is important to note that due to the choice of the possible combinations, the MRF model will not output radically different decisions from the initial classifier. In practice, it makes small changes in the sequence of the initial belief vectors, which is further confirmed by the experimental results in Table 4.

## 6. Experiments and classification results

After having presented promising segmentation results on the test scenes, an extensive analysis of mining object co-occurrences from Flickr data and a description of the proposed MRF, we now present our classification results on the RGB-D and NYU datasets. The experimental analysis of our recognition (binary and multi-class) experiments will be presented in Sections 6.1 and 6.2. An extended set of experiments for our multi-class classifier with MRF context model are reported in Section 6.3. We have also tested our multi-class classifier along with the MRF context model on the NYU scene dataset and we report our numerical analysis of the results in Section 6.4. Our binary and multi-class object classifiers (with and without context modeling) were trained with the RGB-D Object dataset and tested on the RGB-D Scenes dataset [1]. Specifically, we train binary classifiers on two object classes: *bowl* and *coffee mug*. For the multi-class object classifier we train six object classes that are annotated in the RGB-D Scenes Dataset: *bowl, cup, coffee mug, cereal box, flashlight* and *soda can*. This dataset

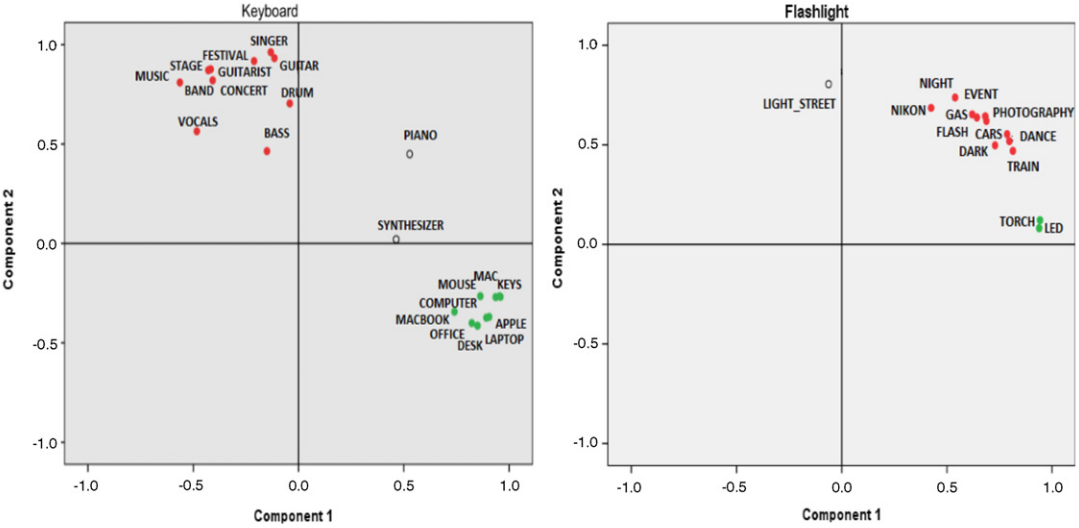


**Fig. 8.** Factor analysis on tags from (left) $\mathcal{D}_s$(keyboard), (right) $\mathcal{D}_s$(flashlight).

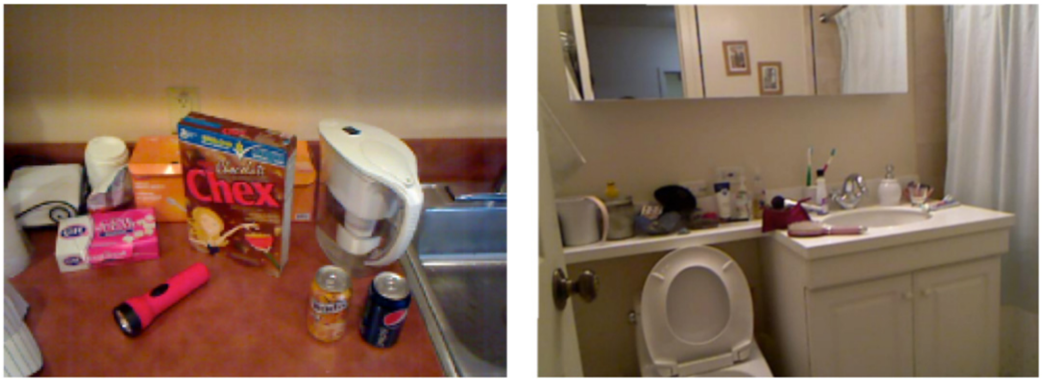


**Fig. 9.** Example of test scenes from the two datasets. RGB-D dataset: kitchen_small_1 (left), NYU dataset V2: bathroom (right).

contains video sequences of scene types, such as: desk, table, meeting room, etc. (Fig. 9). For testing our context-aware classifier approach, we need at least 2 objects in each scene in order to have co-occurrence. Therefore, only those scenes which contain at least 2 annotated objects are considered. Our approach does not try to tackle the problem of object recognition of minimal partial views of objects or objects containing occlusions. We assume an optimal segmentation of the objects in the scenes, therefore annotations of minimal partial views of objects are filtered as well. The filtering steps on the scenes dataset are applied to every 5th frame from the video sequence of each scene type. We chose to sample every 5th frame to directly compare our classification results from the ones presented in [1].

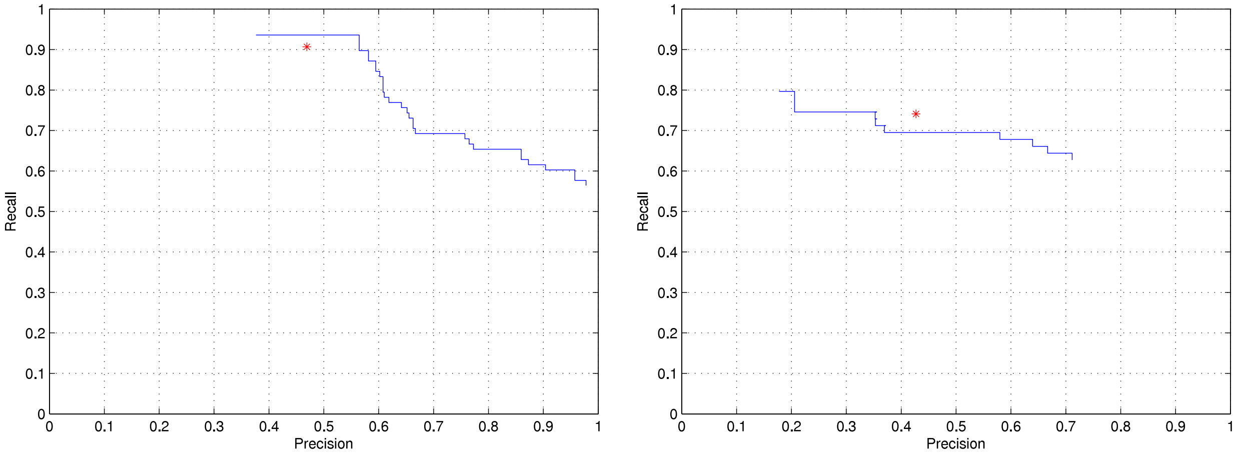
### 6.1. Binary object classification

In this section, we evaluate two binary classifiers and compare the results with the approach presented in [12]. This section provides the evaluation scheme, LR parameter selection and the dataset details. The L2 regularized logistic linear classifier [42] that we have chosen provides predicted labels and confidence estimates at its output. We carried out multiple experiments and derived promising results. What is most important is the relation between the two ROC curves of the classifiers to be compared; and more specifically, their precision/recall values for the chosen point of operation. Given that we do not have an ROC curve from the other approach (i.e. [12]), we decided to plot ROC curves for our classifier, and at least compare our full ROC to a single point from the other classifier. You can see the results in Fig. 10. What becomes evident is that for the case of the object bowl, our classifier is superior at that point of operation, and not significantly inferior for the case of the object coffee_mug. This is evidence that the performance is certainly comparable, and might well be superior. In order to fully verify this claim, though, one would need the full ROC curve from the other method (i.e. [12]), which has not been provided by the other authors.

### 6.2. Multi-class object classification

In this section, we evaluate the multi-class classifiers on the RGB-D scene Dataset. We have trained a multi-class classifier using 6 object categories (bowl, cup, cereal box, coffee mug, flashlight, soda can) from 51 RGB-D Dataset object categories. To produce a probabilistic output to the multi-class problem we use a linear SVM (from LIBSVM [43]). The results from our proposed RGB-D object classifier are presented in Table 2. The accuracy of our classification framework (63.91%) is four times the minimum baseline generated by a random guess (16.67%).

The *# of objects* listed in the table corresponds to the objects of the 6 categories considered in this experiment. The difference in performance between scene types is due to the variations in clutter and viewpoints of the scene. For example, the highest accuracy achieved (83%) was in the *table_1* scene type (Fig. 9, right)—these scenes present similar viewpoints to those used in the object dataset. The object dataset was obtained by recording video sequences of each object as it spun around on a turntable.



(a) ROC for *bowl* binary classifier.

(b) ROC for *coffee_mug* binary classifier.

**Fig. 10.** ROC curves for binary classifier of the following object categories: (a) bowl and (b) coffee_mug. The red point is a precision/recall reported in [12]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

**Table 2**
Accuracies of a 6-category object classifier trained on RGB-D objects dataset and tested on RGB-D scenes dataset. Note that the use of context using MRF significantly improves the accuracy of the baseline classifier.

| Scene type | # of scenes | # of objects | Accuracy (%) | Using Flickr context (%) |
|---|---|---|---|---|
| Desk_1 | 8 | 17 | 64.70 | 70.59 |
| Desk_2 | 2 | 4 | 50.00 | 50.00 |
| Desk_3 | 15 | 36 | 50.00 | 58.33 |
| Kitchen_small_1 | 12 | 31 | 35.48 | 35.48 |
| Table_1 | 18 | 100 | 83.00 | 83.00 |
| Table_small_1 | 19 | 42 | 52.38 | 54.76 |
| *Overall* | **74** | **230** | **63.91** | **66.09** |

**Table 3**
Accuracy results for the NYU dataset vs. random baseline.

| Dataset | Top1 | Top2 | Top3 | Top4 | Top5 | Baseline |
|---|---|---|---|---|---|---|
| NYU6 | 62.22 | 88.89 | 95.56 | 97.78 | 97.78 | 16.67 |
| NYU12 | 53.85 | 76.92 | 86.54 | 88.46 | 88.46 | 8.33 |
| NYU15 | 51.79 | 69.64 | 82.14 | 83.93 | 87.50 | 6.67 |

The data were recorded with the cameras mounted at one meter from the turntable at three different heights relative to it, at approximately 30°, 45° and 60° above the horizon. Thus, on a scene type like *kitchen_small_1* (Fig. 9—left), which presents the lowest classification rate (35.48%) the viewpoint of the camera is higher than 60° above the horizon w.r.t. the table. Furthermore, this scene type presents higher clutter which reduces the performance of the segmentation algorithm and consequently the classifier.

In order to further validate our system, we decided to also train on part of the NYU dataset, and test on the remaining part, before also providing results showing the cross-domain generalization ability of our system (train in RGB-D, test in NYU) in Section 6.4. In Table 3 you can see classification results for training and testing on mutually exclusive subsets of the NYU dataset. Notice that in the NYU dataset scenes that were used, there were 21 total object categories for which we also had co-occurrence information. It is worth noting that many categories have very few instances; for example, 6 out of 21 have only one instance. It is clear that with only one instance, one cannot train and test; so reporting results for this would have been impossible. Furthermore, more than half of the objects have less than 5 instances; and for these, one would need to have a 2–3 object training set, which again would create meaningless results, even more so for the high-dimensional feature representation that we are using. For this reason, if one trains in the NYU dataset and tests on a different part of the same dataset, he would have very few objects with a large enough number of instances in order to be able to train our classifiers appropriately. Therefore, we further restricted the number of objects, in order to have at least 3 instances (for set NYU15, as mentioned in Table 3), or at least 5 instances (for set NYU12) or at least 10 instances (for set NYU6). Even though the number of training pictures is small, as we have achieved overall classification accuracies reaching all the way up to 90% for the case of top-2 rank, the results are indeed good.

### 6.3. *Multi-class object classification using context modeling*

In Table 2 we also compare the results of using our multi-class object classifier combined with an MRF context model extracted from Flickr co-occurrences. Notice that our co-occurrence model captured by the MRF is based on Flickr, i.e. it is not based on our training and testing datasets co-occurrences. As our model was derived from Flickr data, in principle it might have different co-occurrence statistics than RGB-D or NYU. Nevertheless, from our results, it is clearly evident that using a context-aware MRF model for object recognition outperforms the multi-class classifier alone (from 63.91% to 66.09% overall recognition rate), and for some scenes there is a large improvement, such as for *desk_1* and *desk_3*. There exist other scenes however (like *desk_2*, *kitchen_small_1* or *table_1*) for which we see that using a Flickr context-aware MRF does not improve the performance, but it certainly does not decrease it. As these scenes are artificial (i.e. specifically constructed by humans for data collection) some of the co-occurrences are not so natural. For example, *coffee mug* and *cereal box* might have a high co-occurrence probability from the mined Flickr metadata. However, *cereal box* and *flashlight* do not have high co-occurrence in the Flickr metadata, but they appear together frequently in this specific dataset. Thus, in a dataset with more natural co-occurrences the MRF would improve performance even further.

### 6.4. *Cross-domain classification using contextual modeling*

To evaluate the applicability of our system as well as its generalization ability on novel domains, we tested our multi-class object classifier (trained on the RGB-D Object Dataset) on the NYU indoor scene dataset version 2 [2]. This Kinect dataset consists of 1449 RGB-D images containing 26 different scene types, spanning 849 unique object types. This dataset provides per-pixel accurate object segmentation masks for all of the scenes, which are suitable for testing our method. We used a small test set of

**Fig. 11.** Segmented objects from NYU dataset V2 scenes (left) bathroom (right) kitchen.

**Table 4**
Accuracies of a 21-category object classifier trained on RGB-D objects dataset and tested on NYU scenes dataset. Note that the accuracy significantly drops partly due to domain change (see Table 2) and partly due to a higher number of object categories involved in this experiment. However, the MRF context-aware classifier is still able to bring improvements in accuracy.

| Scene type | # of scenes | # of objects | Multi-class classifier (%) | Using Flickr context (%) |
|---|---|---|---|---|
| Bathroom | 7 | 19 | 21.05 | 21.05 |
| Bedroom | 3 | 6 | 16.66 | 16.66 |
| Classroom | 5 | 11 | 9.09 | 9.09 |
| Computer_lab | 6 | 20 | 25.00 | 25.00 |
| Dining_room | 10 | 23 | 8.69 | 8.69 |
| Furniture_store | 12 | 79 | 6.32 | 7.59 |
| Home_office | 4 | 9 | 11.11 | 11.11 |
| Kitchen | 39 | 109 | 10.09 | 11.01 |
| Office | 4 | 8 | 37.50 | 37.50 |
| *Overall* | **90** | **284** | **11.61** | **12.32** |

scenes which contained at least 2 unique objects that correspond with the 51 objects categories from the RGB-D Object Dataset. These scene types contain the following 21 object categories corresponding to the RGB-D Object Dataset: *apple, ball, banana, binder, bowl, calculator, flashlight, keyboard, lemon, notebook, onion, orange, peach, plate, potato, scissors, soda can, sponge, stapler, toothbrush* and *toothpaste*. This experimental setup is a *difficult* task, since we attempt to recognize objects trained from one dataset on a completely different dataset, without considering any domain-adaptation. Thus, for this experiment we expect the multi-class object classifier to perform poorly. As can be seen in Fig. 9, these natural scene types collected from the NYU dataset are quite different than the scenes provided by the RGB-D Dataset. The main differences are: (i) the position of the sensor when the image was taken affecting the scale of the objects in RGB (Fig. 11) as well as in the Depth image which consequently creates different feature vectors and (ii) these scenes are real scenes taken from buildings in three cities, and this produces a larger variety of object category instances which is not seen in the RGB-D object Dataset. Nevertheless, this experiment is aimed at demonstrating the gain of combining context-aware modeling with multi-class object classifiers.

In Table 4 we present the classification results for nine different NYU dataset scene types. As predicted, the Flickr context-aware multi-class object classifier does improve the performance of the initial classifier, however with not as high an improvement as the one achieved in the first case. This is due to two main reasons. Firstly, we are searching for a minimal subset of objects from the wide range of objects that appear in this dataset, namely 21 objects from 849 unique objects in total—merely 2% of the full set of objects. This causes the variation in different objects within the scene to be extremely low, and in a number of the scene types the classifier aided by Flickr-context shows no improvement. But if we analyze one of the scene types in detail, say the *bathroom* scene type, we see that in one scene we have the following objects {*toothbrush*, *toothbrush*, *toothbrush*, *toothbrush*}; in another one we have {*toothbrush*, *toothbrush*, *toothbrush*, *toothpaste*}, and this causes a great difficulty for the MRF model to find object co-occurrences that will boost the results. On the other hand, if we analyze the *kitchen* scene type in detail, here we have a scene with {*apple*, *orange*, *onion*, *banana*, *plate*, *bowl*} and after using the MRF we had a 1% increase from a mere 10% classification result, thus the more natural the co-occurrences in a scene the better the MRF from Flickr data performs. A second reason for the low increase are the low perceptual classification results, prior to the MRF application. For example, if the classifier feeds the MRF with highly erroneous predictions, with the true category being given very low rank and small confidence, it is very difficult for the MRF to improve the accuracy by using context. However, our results illustrate that when having close to natural co-occurrences, classification with contextual modeling in most cases definitely improves performance.

Moving on from scenes to objects, we can observe that the objects which get a more frequent improvement in recognition due to the context model are the following: *plate* was the top object for the NYU dataset, while the top object for the UoW dataset was *bowl* with 5 cases, *soda_can* coming second and *cup* third. Now let us consider how much the classification output is affected by the context model. Out of 284 classifiers decisions, in the NYU dataset, 16 of them (i.e. 5.63%) were changed due to the context modeling, as compared to the multi-class classifier only case. However, not all of these decision changes led to correct recognition; 12.5% of them indeed changed the result so that it is correct, whereas the remaining 77.5% changed the decision, but the new category was still incorrect. In the UoW dataset, 19 out of 230 decisions were affected by context, with the 42.11% of them leading to a correct result. It is worth noting though, that we never had any case of a change from a correct decision to a false decision; i.e. an orange that was thought of as an apple might have been thought of as a banana after the application of the MRF context bias; however, it was never the case that an orange that was thought of as an orange ended up being thought of as something else. Thus, at least in our case, application of the contextual model never deteriorates the results, but only improves them. Finally, it is worth noting, as we mentioned before, that the NYU dataset is a considerably more *difficult* dataset than RGB-D—and thus this clearly illustrates the power of the co-occurrence based MRF model for such a case.

## 7. Discussion and future work

We have started our results by illustrating the convergence of the co-occurrence statistics obtained by mining Flickr. However, it is worth noting that there are at least four different target distributions that appear in the bigger picture of our problem setting. First, there is the real co-occurrence distributions for objects in the earth at large. Second, there is the co-occurrence distribution for labeled objects in Flickr. And third, and fourth, correspondingly, there are the co-occurrence distributions for the training and testing sets in question—in our case, these are coming from sampling the RGB-D or NYU datasets. Apart from these four nominal distributions, there also exist estimates of the above four, which are created when statistics are calculated through subsets (samples) of the above four sets. By illustrating convergence in Section 4, we have shown that one can adequately estimate the second distribution (co-occurrence in Flickr) given subsets of workable size, on the order of 3000 images or so per term. However, this does not guarantee that our estimated co-occurrence distribution of Flickr converges to the real co-occurrence distribution of objects in the world (the first distribution); and it also does not guarantee that it converges to the specific distributions that exist in the training and testing sets under consideration (the third and fourth distributions).

Nevertheless, the distributions obtained are adequate for performance improvement, and as a future step, one could envision using them as an initial estimate which is updated online with the co-occurrence patterns of the specific environment that the robot navigates in, so that it can slowly converge to the actual distribution. As mentioned before, even the distributions obtained from Flickr alone, as our results illustrate, certainly never reduce performance—in most cases they contribute to improved performance.

Most importantly, all the results presented in the paper, with or without MRF context models, are significantly higher than the previous results reported regarding the RGB-D database, such as [1]. Thus, in short, our novel object recognition method has improved performance and, most importantly, can accommodate further improvements using an appropriately sampled co-occurrence distribution in order to create an MRF combined with our classifiers, producing highly noticeable results, even in difficult datasets.

Regarding other future steps, we are currently improving our segmentation methods, in order to solve the discontinuity and hole problems that we have sometimes faced. Furthermore, we are investigating the usage of complete reconstruction of partial-view 3D models accumulated across multiple viewpoints, as they exist within the camera movement trajectories. Most importantly, we plan to investigate the question of the relation of recognition confidence as a function of the number and type of views; ultimately towards creating an active vision model which steers the camera trajectory in order to minimize uncertainty with minimum movement and time cost. Given that robots and other devices nowadays usually have some mobility capability, it is highly worthwhile to utilize this in order to derive even better results.

## 8. Conclusion

In this paper, we have investigated the use of a new object recognizer as well as an object context model driving a Markov Random Field towards object recognition in RGB-D images. The co-occurrence probabilities for the object context model were derived from our datasets as well as from the Flickr online database. After introducing relevant existing literature, we started by investigating methods and convergence properties for the co-occurrence distributions. Then, after a description of our system and experiments, we presented multiple quantitative results as well as comments and a discussion.

The results are indeed good: our novel combination of viewpoint histogram features for shape description with standard color features yielded competitive accuracy for recognizing categories of isolated objects. Our Markov Random Field based model for utilizing context information to make more informed decisions about the possible category of target objects in a cluttered scene provided further improvements. We showed that social media mining can be effectively used for computing co-occurrence probabilities of objects in natural scenes, for example in the case of the very difficult NYU dataset. The first set of experiments with the publicly available RGB-D Dataset show interesting insights into the use of contextual modeling for object category recognition. The second set of cross-domain experiments (training the classifier on RGB-D Dataset and testing on NYU dataset) demonstrates that a significant accuracy gain can be achieved by utilizing context information in novel domains.

In conclusion, we believe that contextual models are an important step towards improved real-world object recognition, and that better understanding of the properties of such models, as well as of the sources for deriving them, is a prerequisite towards their successful application. Through our work, we have illustrated how such models can be created and used, so that in conjunction with cost-effective depth cameras, such as the Kinect, we can get closer to the ultimate goal of widely available intelligent devices with real-world scene understanding capabilities.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.robot.2013.10.001.

## References

[1] K. Lai, L. Bo, X. Ren, D. Fox, A large-scale hierarchical multi-view RGB-D object dataset, in: Robotics and Automation, ICRA, Shanghai, China, pp. 1817–1824.

[2] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from RGBD images, in: 2012 European Conference on Computer Vision, ECCV, Florence, Italy, pp. 746–760.

[3] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110.

[4] H. Bay, A. Ess, T. Tuytelaars, L. van Gool, SURF: speeded up robust features, Computer Vision and Image Understanding 110 (2008) 346–359.

[5] P. Henry, M. Krainin, E. Herbst, X. Ren, D. Fox, RGB-D mapping: using kinect-style depth cameras for dense 3D modeling of indoor environments, International Journal of Robotics Research 31 (2012) 647–663.

[6] X. Ren, L. Bo, D. Fox, RGB-(D) scene labeling: features and algorithms, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, CVPR 2012, 2012, pp. 2759–2766.

[7] R. Rusu, S. Cousins, 3D is here: point cloud library, in: Robotics and Automation, ICRA, 2011.

[8] L.A. Alexandre, 3D descriptors for object and category recognition: a comparative evaluation, in: Workshop on Color-Depth Camera Fusion in Robotics at the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vilamoura, Portugal.

[9] F. Tombari, S. Salti, L. Di Stefano, A combined texture-shape descriptor for enhanced 3D feature matching, in: 2011 18th IEEE International Conference on Image Processing, ICIP, pp. 809–812.

[10] N. Figueroa, H. Ali, F. Schmidt, 3D registration for verification of humanoid Justin's upper body kinematics, in: 2012 Ninth Conference on Computer and Robot Vision, CRV, pp. 276–283.

[11] M. Blum, J.T. Springenberg, J. Wülfing, M. Riedmiller, On the applicability of unsupervised feature learning for object recognition in RGB-D data, in: Workshop on Deep Learning and Unsupervised Feature Learning at the NIPS 2011, Granada, Spain.

[12] K. Lai, L. Bo, X. Ren, D. Fox, Detection-based object labeling in 3D scenes, in: Robotics and Automation, ICRA, 2012, pp. 1330–1337.

[13] L. Bo, X. Ren, D. Fox, Unsupervised feature learning for RGB-D based object recognition, in: Experimental Robotics, ISER, 2012.

[14] H. Ali, F. Shafait, E. Giannakidou, A. Vakali, N. Figueroa, T. Varvadoukas, N. Mavridis, Addendum to contextual object category recognition for RGB-D scene labeling, 2013. http://oswinds2.csd.auth.gr/~irini/.

[15] M. Blum, J.T. Springenberg, J. Wulfing, M. Riedmiller, A learned feature descriptor for object recognition in RGB-D data, in: Robotics and Automation, 2012, ICRA, IEEE, 2012, pp. 1298–1303.

[16] E. Nascimento, G. Oliveira, M. Campos, A. Vieira, W. Schwartz, Brand: a robust appearance and depth descriptor for RGB-D images, in: Intelligent Robots and Systems, IROS, 2012, pp. 1720–1726.

[17] R.B. Rusu, N. Blodow, Z.C. Marton, M. Beetz, Aligning point cloud views using persistent feature histograms, in: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, pp. 3384–3391.

[18] A. Zaharescu, E. Boyer, K. Varanasi, R. Horaud, Surface feature detection and description with applications to mesh matching, in: Computer Vision and Pattern Recognition, 2009, CVPR 2009, pp. 373–380.

[19] A. Kanezaki, Z.-C. Marton, D. Pangercic, T. Harada, Y. Kuniyoshi, M. Beetz, Voxelized shape and color histograms for RGB-D, in: Intelligent Robots and Systems, IROS, 2011, Workshop on Active Semantic Perception and Object Search in the Real World, San Francisco.

[20] R.B. Rusu, G.R. Bradski, R. Thibaux, J. Hsu, Fast 3D recognition and pose using the viewpoint feature histogram, in: IEEE Intelligent Robots and Systems, IROS, 2010, pp. 2155–2162.

[21] T. Malisiewicz, A. Efros, Recognition by association via learning per-exemplar distances, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, CVPR 2008, pp. 1–8.

[22] S. Golder, B. Huberman, The structure of collaborative tagging systems, Journal of Information Science 32 (2005) 198–208.

[23] C. Marlow, M. Naaman, D. Boyd, M. Davis, HT06, tagging paper, taxonomy, Flickr, academic article, to read, in: Hypertext and Hypermedia, Odense, Denmark, 2006, pp. 31–40.

[24] X. Olivares, M. Ciaramita, R. van Zwol, Boosting image retrieval through aggregating search results based on visual annotations, in: Multimedia, Vancouver, British Columbia, Canada, 2008, pp. 189–198.

[25] L. Kennedy, M. Naaman, S. Ahern, R. Nair, T. Rattenbury, How Flickr helps us make sense of the world: context and content in community-contributed media collections, in: Multimedia, 2007, pp. 631–640.
[26] D. Crandall, L. Backstrom, D. Huttenlocher, J. Kleinberg, Mapping the World's photos, in: World Wide Web, Madrid, 2009, pp. 761–770.
[27] T. Rattenbury, N. Good, M. Naaman, Towards automatic extraction of event and place semantics from Flickr tags, in: Research and Development in Information Retrieval, Amsterdam, 2007, pp. 103–110.
[28] T. Quack, B. Leibe, L. Van Gool, World-scale mining of objects and events from community photo collections, in: Content-based Image and Video Retrieval, Niagara Falls, Canada, 2008, pp. 47–56.
[29] G. Begelman, P. Keller, F. Smadja, Automated tag clustering: improving search and exploration in the tag space, in: WWW'06 Conference on Collaborative Web Tagging Workshop, Edinburgh, UK, 2008, pp. 22–26.
[30] F. Eggenberger, G. Polya, Uber die statistik verketter vorgage (the statistics of chained operations), ZAMM—Journal of Applied Mathematics and Mechanics 1 (1923) 279–289.
[31] H. Halpin, V. Robu, H. Shepherd, The complex dynamics of collaborative tagging, in: World Wide Web, Banff, 2007, pp. 211–220.
[32] A. Nüchter, J. Hertzberg, Towards semantic maps for mobile robots, Robotics and Autonomous Systems 56 (2008) 915–926.
[33] X. Xiong, D. Huber, Using context to create semantic 3D models of indoor environments, in: Proceedings of the British Machine Vision Conference, BMVA Press, 2010, pp. 45.1–45.11.
[34] H. Koppula, A. Anand, T. Joachims, A. Saxena, Semantic labeling of 3D point clouds for indoor scenes, in: Conference on Neural Information Processing Systems, NIPS, 2011, pp. 244–252.
[35] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, International Journal of Computer Vision 42 (2001) 145–175.
[36] A. Oliva, A. Torralba, et al., The role of context in object recognition, Trends in Cognitive Sciences 11 (2007) 520–527.
[37] C. Galleguillos, A. Rabinovich, S. Belongie, Object categorization using co-occurrence, location and appearance, in: CVPR 2008, IEEE, 2008, pp. 1–8.
[38] T. Kollar, N. Roy, Utilizing object–object and object-scene context when planning to find things, in: ICRA, 2009, IEEE, 2009, pp. 2168–2173.
[39] Y. Jiang, C. Ngo, S. Chang, Semantic context transfer across heterogeneous sources for domain adaptive video search, in: Multimedia, Beijing, China, 2009, pp. 155–164.
[40] E. Chatzilari, S. Nikolopoulos, I. Kompatsiaris, E. Giannakidou, A. Vakali, Leveraging social media for training object detectors, in: Digital Signal Processing, Santorini, Greece, 2009, pp. 232–239.
[41] T. Leung, J. Malik, Representing and recognizing the visual appearance of materials using three-dimensional Textons, International Journal of Computer Vision 43 (2001) 29–44.
[42] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: a library for large linear classification, Journal of Machine Learning Research 9 (2008) 1871–1874.
[43] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 27:1–27:27.
[44] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM 24 (1981) 381–395.
[45] R.B. Rusu, Semantic 3D object maps for everyday manipulation in human living environments, Ph.D. Thesis, Computer Science department, Technische Universitaet Muenchen, Germany, 2009.
[46] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: a comprehensive study, International Journal of Computer Vision 73 (2007) 213–238.
[47] A. Janoch, S. Karayev, Y. Jia, J. Barron, M. Fritz, K. Saenko, T. Darrell, A category-level 3-D object dataset: putting the kinect to work, in: Computer Vision Workshops, ICCV Workshops, pp. 1168–1174.
[48] C. Cortes, L. Jackel, S. Solla, V. Vapnik, S. Denker, Learning curves: asymptotic values and rate of convergence, in: Neural Information Processing Systems, NIPS, Denver, Colorado, 1993, pp. 327–334.
[49] P. Billingsley, Convergence of Probability Measures, second ed., John Wiley and Sons, USA, 1999.
[50] N. Breslow, A generalized Kruskal–Wallis test for comparing $k$ samples to unequal patterns of censorship, Biometrika 57 (1970) 579–594.
[51] E. Giannakidou, V. Koutsonikola, A. Vakali, I. Kompatsiaris, Co-clustering tags and social data sources, in: Web-Age Information Management, Zhangjiajie, China, 2008, pp. 317–324.
[52] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.

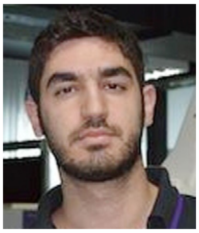
**Haider Ali** is currently serving as Senior Researcher at the Institute of Robotics and Mechatronics (RM) of the German Aerospace Center (DLR). His research is focused on developing efficient 3D object detection and pose estimation methods for real time robotics applications. He received his Bachelor of Science in Computer Science from Bahauddin Zakariya University one of Pakistan's major universities in 1998. After that he served several multinational IT companies in Pakistan as Software Engineer and Project Consultant until 2004. Thereafter he planned to pursue a master degree in software technology from Leuphana University of Lueneburg and graduated in 2006. He received his Ph.D. from the Technical University of Vienna in 2010.

**Faisal Shafait** is working as a Research Assistant Professor in the Computer Science & Software Engineering Department at the University of Western Australia. Formerly he was a Senior Researcher in the Multimedia Analysis and Data Mining (MADM) competence center at the German Research Center for Artificial Intelligence (DFKI) as well as an Adjunct Lecturer at Kaiserslautern University of Technology (TUKL), Germany. He received his Ph.D. with the highest distinction in computer engineering from TUKL in 2008. His research interests include machine learning and pattern recognition with a special emphasis on applications in document image analysis. He has co-authored over 80 publications in international peer-reviewed conferences and journals in this area. He is an Editorial Board member of the International Journal on Document Analysis and Recognition (IJDAR), a Technical Program Committee member of the 12th International Conference on Document Analysis and Recognition (ICDAR) held in Washington DC, USA in Aug. 2013 as well as the 5th International Workshop on Computational Forensics, held in Tsukba, Japan in Nov. 2012. He was a PC Chair of the 4th International Workshop on Camera-Based Document Analysis and Recognition, held in Beijing, China in September 2011.

**Eirini Giannakidou** is a Ph.D. candidate at Aristotle University of Thessaloniki and a researcher in Informatics and Telematics Institute, Greece. Her research focuses on social tagging and its potential impact on improving web information retrieval. She received a B.Sc. in Computer Science and an M.Sc. in Information Systems from Aristotle University of Thessaloniki. Contact her at eirgiann@csd.auth.gr.

**Athena Vakali** has been a faculty member (now an associate professor) in the Department of Informatics at the Aristotle University of Thessaloniki since 1997. She is heading the Operating Systems Web/INternet Data Sources Management research group "OSWINDS" (http://oswinds.csd.auth.gr). Her research activities are on various aspects and topics of the Web information systems, including Web data management (clustering techniques), content delivery on the Web, Web data clustering, Web caching, XML-based authorization models, text mining and multimedia data management. Her publication record is now at more than 100 research publications which have appeared in several journals (e.g CACM, IEEE Internet Computing, WWWJ), book chapters and in scientific conferences (e.g., IDEAS, ADBIS, ISCIS, ISMIS etc.). In March 2004, she co-organized the EDBT-Workshop on Clustering Information over the Web (ClustWeb), in the IX Conference on Extending Database Technology (EDBT) in Heraclion, Greece. In April 2005 and April 2006, she co-organized the ICDEWorkshops on Challenges in Web Information Retrieval and Integration (WIRI), in Tokyo (Japan) and in Atlanta (USA) respectively. She is regular reviewer in major Web data management conferences (i.e., ECML/PKDD, EDBT, CoopIS, DASFAA) and journals (i.e., IEEE Internet Computing, IEEE TKDE, DKE). She is also co-editor of the book "Web Data Management Practices: Emerging Techniques and Technologies" published by Idea Group Publishing. She is a member of the editorial board of the Computers and Electrical Engineering Journal (Elsevier) and since March 2007, she is the coordinator of the IEEE TCSC technical area of Content Management and Delivery Networks.

**Nadia Figueroa** received her M.Sc. in Automation and Robotics from the Technical University of Dortmund in 2012, after receiving her B.Sc. in Mechatronics from Monterrey Tech (Mexico). During her studies in Germany, she worked on a number of research topics such as automated vehicles, object recognition, kinematic calibration, as well as machine vision. Her thesis entitled: "3D Registration for Verification of Humanoid Justin's Upper Body Kinematic", developed at the Institute of Robotics and Mechatronics (RM) of the German Aerospace Center (DLR) addressed the problem of self-verification of the complex kinematic chains of the humanoid robot "Rollin' Justin" by using external sensory systems such as: multi-camera IR tracking system or the On-Board Stereo Camera system. She is currently a research assistant at the Engineering Department of New York University Abu Dhabi (NYUAD). Her research interests include: robotics, computer vision, artificial intelligence and cognitive systems.



**Theodoros Varvadoukas** received his B.Sc. in Informatics and Telecommunications from the National and Kapodestrian University of Athens in 2012. The research during his thesis was conducted at the National Center for Scientific Research (NCSR) Demokritos. His work entitled "Detecting Human Patterns in Laser Range Data" resulted in a publication at the ECAI '12, where a novel pattern recognition-based approach was proposed for the problem of human detection from a 2D laser sensor, avoiding the need for predefined motion models. During his thesis he was involved in several research modules on the Roboskel robotic platform such as NLP, motion planning, SLAM, as well as in classic machine learning and pattern recognition problems. He is currently a research assistant at New York University Abu Dhabi (NYUAD). His research interests include: interactive robotics, machine learning & vision and cognitive systems.

**Nikolaos Mavridis** received his Ph.D. from MIT in 2007, after receiving his M.S.E.E. from UCLA and a M.Eng. in ECE from the Aristotle University of Thessaloniki. From 2007 to 2011, he served as Assistant Professor at the College of IT of the United Arab Emirates University, where he founded the Interactive Robots and Media Laboratory (IRML). He is currently serving as Assistant Professor at New York University AD, where his lab is now located. The lab is home to the "FaceBots" social robots project, as well as to "IbnSina", the first Arabic-language conversational Android robot. In his Ph.D. Thesis at MIT, he introduced the "Grounded Situation Model" proposal, and demonstrated its benefits by implementing it on Ripley, a manipulator robot with vision, touch, and speech synthesis/recognition. The sensory motor/linguistic abilities of the resulting system were comparable to those implied by a standard psychological test for 3-year old children (The "Token Test"). The research interests of Dr. Mavridis include Robotics, especially Interactive and Social Robotics, as well as Cognitive Systems.