2012 10th International Conference on Frontiers of Information Technology

Indoor Furniture and Room Recognition for a Robot using Internet-derived Models and Object Context

Theodoros Varvadoukas[1], Eirini Giannakidou[2], Javier V. Gómez[3] and Nikolaos Mavridis[4]

***Abstract*— For robots to be able to fluidly collaborate with and keep company to humans in indoor spaces, they need to be able to perceive and understand such environments, including furniture and rooms. Towards that goal, we present a system for indoor furniture and room recognition for robots, which has two significant novelties: it utilizes internet-derived as well as self-captured models for training, and also uses object- and room-context information mined through the internet, in order to bootstrap and enhance its performance. Thus, the system also acts as an example of how autonomous robot entities can benefit from utilizing online information and services. Many interesting subproblems, including the peculiarities of utilizing such online sources, are discussed, followed by a real-world empirical evaluation of the system, which shows highly promising results.**

I. INTRODUCTION

In every day life humans wander around buildings interiors. Those spaces contain complex environments, which, however, show some regularity that can be formalized. For an harmonic symbiosis between humans and interactive robots to emerge, the latter should develop good enough skills for indoor scene understanding. Furnitures and rooms are structural elements of every building interior. For that reason, robots need to acquire capabilities for recognizing and estimating different pieces of furniture and different types of rooms, so that they achieve fluid collaboration and communication with humans.

Regarding a relevant yet much more generic problem, namely object recognition, there has been quite some research for the last two decades. Nowadays, the interest in object recognition has been reinforced with the evolution of the hardware equipment, especially after the release of the relatively low-cost Prime Sense RGB-Depth sensors in the Microsoft Kinect camera or in the ASUS Xtion Pro LIVE. These devices provide the typical RGB planar image but also capture depth information for every pixel with VGA resolution (640x480 pixels) at approximately 24 Hz. This information is very helpful for any object segmentation and recognition tasks.

On the other hand, recent research is focusing on both taking into account object context and using World Wide Web data (such open datasets or 3D CAD models) to extract knowledge and augment the training data. More specifically, in most social media websites, users are encouraged to provide a short description, in terms of tags or keywords, for any kind of resource they create. Research has shown that these descriptions are not randomly chosen, but quite a large portion of them rather summarizes the main topic of each resource [1].

In this paper, we develop a 3D object classifier trained and tested upon different mixtures of web 3D models and real-world data. We then extract automatically terms related to indoor environments using the online database *Wordnet* and query them as tags to *Flickr* in order to retrieve co- occurrence probabilities. Both sources of information, classifier's class-wise accuracy and co-occurrence probabilities, are then combined and create a *Markov Random Field* (MRF) model which consist our final probabilistic classifier. We further exploit the MRF's structure and contextual information in order to obtain a belief about any novel, for the 3D classifier, object as well as for the surrounding environment.

The paper is organized as follows: Section II provides an extensive background in object context and recognition. Section III describes the architecture we propose for 3D object recognition taking into account co-occurrence probabilities taken from the World Wide Web. In Section 4 the preliminary results are shown. A discussion is carried out in Section V. Finally, in Section VI the main conclusions of the paper are outlined.

II. BACKGROUND

A. Robotic Vision for Indoor Environments

The design of a robotic vision system is highly influenced by the field of application of such systems. One of the most important aspects to take into account is if the robot is going to operate outdoor or indoor.

In outdoor environments, there are often, not very cluttered, large free spaces. However, the moving speed of the robot is generally higher than indoor environments as in the case of autonomous cars. In those cases, Kinect-like cameras are not useful due to their low range (maximum of 7 meters), small field of view and their working principle, based on infrared light (IR). Hence, 3D laser range scanners such as the Velodyne are very well suited for these applications.

However, indoor scenarios require more detail to be described, since they usually are very cluttered. In the case of object recognition, Kinect-like cameras or stereoscopic systems are the best solution, since they provide high-density

[1] T. Varvadoukas is with Institute of Informatics and Telecommunications, NCSR Demokritos and Department of Informatics, University of Athens, Greece t.varvadoukas@gmail.com

[2] E. Giannakidou is with Informatics Department, Aristotle University of Thessaloniki, Greece eirgiann@csd.auth.gr

[3] J.V. Gómez is with RoboticsLab, Carlos III University of Madrid, Av. de la Universidad 30, 28911, Madrid, Spain jvgomez@ing.uc3m.es

[4] N. Mavridis is with Department of Computer Engineering, New York University Abu Dhabi, Abu Dhabi, UAE nikolaos.mavridis@nyu.edu

978-0-7695-4927-9/12 $26.00 © 2012 IEEE
DOI 10.1109/FIT.2012.30

point clouds and also RGB information. More over, Kinect-cameras are able to obtain 3D point clouds even without any light source. This is impossible for stereoscopic cameras. The main limitations of these cameras are not important in indoor environments, since the speed of the robot is going to be slow and the range is enough for most of the buildings' interiors. Recent works in indoor environment mapping are based on these devices [2], [3].

Focusing on assistive robotics, our main goal are indoor environments. Due to this fact, we choose as sensor a Kinect camera. However, we do not take into account the RGB information, since one of the objectives of this work is to create a framework which also works with other kind of sensors, such as laser range scanners.

B. 3D Object Recognition

Since object recognition is a very generic category, we focus on 3D object recognition. The methods for 3D objects retrieval can be classified into four main groups, according to [4]: *Histogram-based*, *Transform-based*, *Graph-based*, *View-based*.

Apart of the aforementioned, there are also combinations of the different groups. An example of combined descriptor with good results is the Viewpoint Features Histogram (VFH) [5]. The VFH combines information about the 3D shape of the object by encoding the relations between the angles of the surface normals for couples of points into a histogram. Even more, the angle between those normals and the viewpoint vector is also computed and attached to the histogram. Therefore, VFH contains information of both view and geometrical shape.

In recent years, multi-domain object recognition, in which objects acquired through many different media, sensors, 3D CAD models or hand drawn sketches, has attracted the community's attention. This provokes the usage of large, on-line databases to train the system without requiring human annotation.

In [6], although applied to outdoor environments, the authors explain a way to train an object classifier by using Google's 3D Warehouse, exploiting datasets available on the World Wide Web. A domain adaptation step was carried out by means of *feature augmentation* [7] which increases the accuracy of the classifier. More specifically, it is based on creating a stacked feature vector which automatically places the objects from the same domain closer than objects from other domains. With the aid of this adaptation, the accuracy of the classifier is increased when trying to identify real objects by using CAD models in the training process.

We are going to use the algorithm proposed in [4] since it supports multi-modal queries, where the object to identify can be a 3D CAD model, a 3D partial view or even a 2D image. The steps of the algorithm are:

- Pose estimation procedure. The scale and position of the object is normalized to lie within a bounding sphere of radius 1 with the center of mass at the center of the coordinate system. For the orientation estimation, a combination of Principal Component Analysis [8] and Visual Contact Area [9] is used. The one which produces the smaller bounding box is then chosen, since generally a smaller bounding box means that the orientation is better estimated.
- A set of uniformly distributed views is taken from every one of the 18 vertices of a regular 32-hedron. Every view has a binary image, which contains only silhouettes, and depth images, where the pixel intensities are proportional to the distance of the 3D object.
- 2D functionals are computed for each view. Concretely the 2D Polar-Fourier Transform, the 2D Zernike Moments and the 2D Krawtchouk Moments. All the descriptors for a single image are put together in one feature vector.
- The matching is carried out by summing up all the distances of the features vector. The query object is then classified into the class which is closer.

The advantages of this method are manifold. It is robust with respect to the objects' level of detail, it provides an unified framework for multi-modal queries and its discriminative power has been shown to be high.

C. Public On-line Sources for Object Recognition

Using public on-line image datasets to facilitate tasks such as object detection and scene recognition has been met in a number of approaches over the past few years ([10]-[12]). The general idea in these approaches is matching the input image with similar images from public datasets. Indicative datasets used for this purpose are the ones in 80 Million Tiny Images [13] and ImageNet [14].

Recently, large on-line repositories of 3D data such as Google 3D Warehouse have launched [15]. Such resources together with the advent of RGB-D cameras have turned a part of scientific community towards studying recovery and representation of 3D geometry information of real world objects ([16]-[18]). As 2D approaches based on retrieving similar images have proven their value in scene interpretation ([10], [13], [19], [20]), it can be inferred that similar techniques based on geometric features could be equally effective for 3D scene interpretation tasks. In fact, the motivation for such techniques is the same for 3D models as for images: the sizes, shapes, orientations, locations and co-occurrence of objects in real world environments are not arbitrary, but rather constrained in ways that can be represented given enough data.

A number of approaches in the graphics community utilize data from online repositories such as Google 3D Warehouse, in an effort to perceive and model how objects are typically arranged in homes ([21], [22]). Moreover, the vision community has begun using 3D Warehouse data to learn about the geometric properties of objects [23]. Additionally this data have been used to facilitate the classification task for 3D to 3D matching with laser scans [6].

D. Context in Object Recognition

Context powerfully influences how humans act and understand things. However, whilst contextual objects and

relationships facilitate object perceptions for humans, in computer vision co-occurrence relationships or, in general, contextual information have not yet been fully exploited and there are algorithms in which contextual objects serve rather as nuisances that may worsen the performance of an object detection task.

The idea of using context relies on the fact that certain objects typically occur in specific environments or are likely to be near or in specific relation to other objects. So, a natural way to represent the context of an object is to describe relationships with other objects. There is a number of research efforts in computer vision that explore how contextual relationships may improve object search and retrieval efficacy, especially in indoor environments which is relevant to the topic discussed in this paper ([24], [25], [3]). Most of these approaches rely on handcoded statements describing relations or constraints, e.g. "Sea is under Sky" or "A Sofa is on top of the Floor". An example of such technique is presented in [24], where spatial information about the environment and conditions between entities are represented using Horn clauses. However the performance of such approaches decreases a lot in case of noisy data and unknown environments.

An alternate way to utilize contextual cues in vision problems is to exploit local features statistics, in order to identify real world scenes (global context) and then place attention on specific scene areas in which the object is most likely to be found ([26], [27]). An empirical study of context in object detection is presented in [28], where the authors attempt a classification of context sources (i.e. semantic, geographic, temporal, cultural, 3D Geometric context, etc)

Recently, there has been an increasing interest by the research community towards approaches that utilize social data information as context for object search and retrieval tasks. Specifically, a number of works have addressed the problem of identifying photos from social tagging systems that depict a certain object, location or event ([29]-[31]). In [29] they analyze location and time information from geotagged photos from Flickr, in order to track tags that have place semantics (i.e. they refer to an object in a restricted location) or event semantics (i.e. they are met in specified time periods). Then, they employ tag-based clustering on these specific tags, followed by visual clustering, in order to capture distinct viewpoints of the object of interest. The same authors in [32] combine tags with content analysis techniques, in order to get groups of music events photos. Likewise, in ([30], [31]) the authors use various modalities of photos (i.e. visual, textual, spatial, temporal proximity), in order to get photo collections in an unsupervised fashion. Another approach towards this direction, that deploys the visual annotations, also known as "notes" in Flickr is described in [33], where it is shown that the image retrieval may improve significantly by combining tags and visual analysis techniques. Apart from the obvious retrieval application, the outcome of these methods can be used for training of concept detection algorithms, as shown in ([34], [35]). The advantage of using social sites like Flickr is that we can obtain a high number of contextual relationships without spending much effort or time. Consequently, as opposed to supervised approaches, there is no limitation on the types of objects that can be trained, since social sites accommodate images depicting a huge variety of objects.

III. SYSTEM ARCHITECTURE

In order to confront with the problem of object and room recognition in indoor environments we intersect and exploit three different sources of information: object and scene co-occurrence probabilities extracted from social media (cf. III-C), real world Microsof Kinect data and 3D CAD models from the web (cf. III-A). In subsection III-B it is described the construction of a 3D object classifier without using color information, then follows the extraction of co-occurrence probabilities (subsection III-C) and finally we combine them in order to build a *Markov Random Field* model and a more accurate probabilistic classifier (subsection III-D). Figure 1 depicts the structure of the whole system architecture.

A. 3D CAD models and Kinect data acquisition

1) Self-recorded Data Set: The self-recorded dataset was acquired using a Microsoft Kinect camera. We placed the objects in front of the camera without physical proximity with other objects for more accurate segmentation and rotated them in order to get several viewpoints of the same object. For each object we got between 10 and 15 shots, depending on its size and symmetrical complexity.

Our classifier is limited to recognize 4 different classes of objects: *sofa*, *chair*, *table* and *cupboard*. [1] Each class consists of a set of different viewpoints shots from 3-4 different objects, namely a set of partial 3D views without RGB information. Examples of the views for one of the chairs is shown in Figure 2.

[1] In the general class *cupboard* we include similar objects like bookcases.

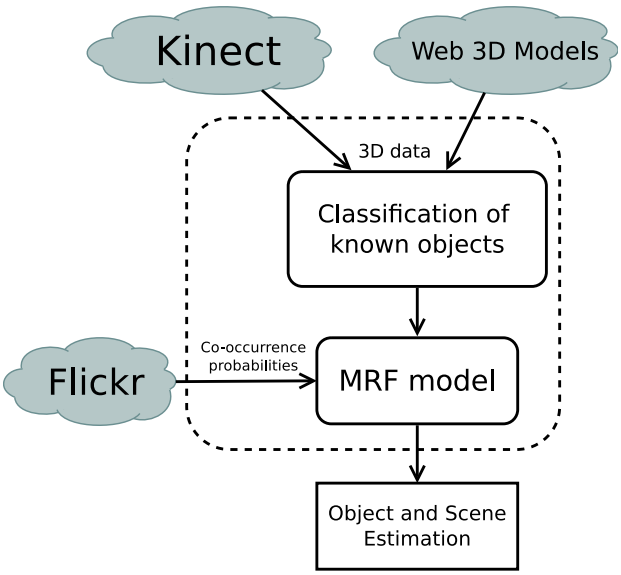

Fig. 1. Schema of the proposed sytem architecture.

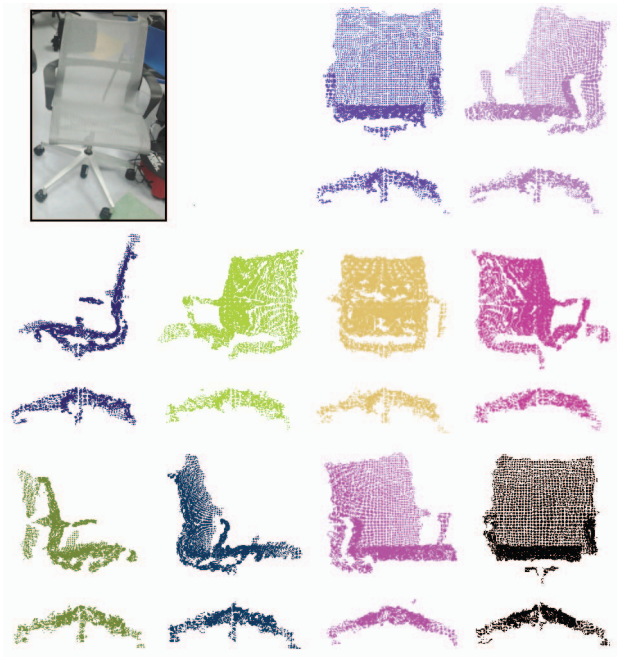

Fig. 2. Picture of the real chair and the 10 different views obtained from it.

2) Online 3D models: The 3D CAD models were downloaded from an online database [2] by querying the 4 types of objects that the classifier recognizes. Figure 3 shows some examples of 3D chairs acquired from the above source. The size of the dataset varies depending on the experiment executed (cf. Section IV for more details).

B. 3D object classification

Our classifier is based on a classical pipeline of segmentation (in case of real-world data), features extraction and categorization using *Euclidean distance* and a *nearest neighbour* approach.

Due to the different character of the data, the classifier should take into account the artificial nature of the online 3D models as well as the lack of information, due to occlusions,

[2]http://archive3d.net, but any other database can be used.

Fig. 3. Three of the 3D CAD models of chairs obtained from the Web.

for the Kinect-based data. A set of normalization steps are essential for a fair comparison. Our method uses the feature extraction algorithm in [4] described in more details in Section II-B. For each 3D model set of 18 distributed uniformly views were sampled. The output of the algorithm is a 156-length feature vector. Finally, each object is classified according to its closest neighbour.

C. Utilizing object/object and object/scene context

This work concentrates on social media and their potential to serve as training sources in an object detection scheme. Social sites like Flickr, accommodate image corpora that are being populated with hundreds of user tagged images on a daily basis. We are interested on whether such corpora can be leveraged to learn contextual relationships among objects and incorporate them into an object detector.

Specifically, here, we focus on Flickr and we treat each image as a vector of tags. Representing images in a latent semantic space captures the correlation between tags and allows hidden relationships to emerge; for example, if $tagA$ and $tagB$ co-occur in a large portion of images they can be mapped to a common latent dimension. We track such latent pairwise relationships by finding matching tag pairs within image annotations and extract tag co-occurrence statistics. We utilize these statistics to predict the likelihood of observing an object o_i given the object o_j, as follows:

$$P(o_i \mid o_j) = \frac{img_{o_i,o_j}}{\sum_{o_k} img_{o_i,o_k}} \quad (1)$$

where the numerator expresses the number of images that tags i, j co-occur and the denominator equals the appearance frequency of tag i with any other object category.

As a source of semantic information for deciding object categories pertaining to indoor environments automatically, we employ the lexicon WordNet. WordNet stores English words organized in hierarchies, depending on their cognitive meaning, and we utilize it to get in an automatic fashion English terms referring to rooms, objects and furnitures. In order for the contextual information to be useful, the extracted co-occurrences and terms should be a good representation of different indoor environments. This can be confirmed in Section IV, where we compare with the co-occurrences of a large and challenging dataset.

D. Probabilistic classifier

The last step of our method is to combine the co-occurence probabilities from social media sources with the 3D object classifier's accuracy in order to create a *Markov Random Field* (MRF) model similar to the one in [36] for enhanced object recognition as well as novel object and scene estimation.

More specifically given a set of M segments s_j extracted from the scene and N classifier's detections on those segments c_i, where $M \geq N$, we want to maximize the following term:

$$P(s_j|c_1, c_2, ..., c_N) = \frac{1}{Z} \prod_{j,k \in M} \psi(s_j, s_k) \prod_{j \in M, i \in N} \phi(s_j, c_i) \quad (2)$$

where ψ is the likelihood of two objects co-occurring, ϕ the classifier's class-wise accuracy and Z a normalization term. An important consequence that arises from formula 2 is that one of the s_j can be a *novel object*, that is an object outside of the fixed number of classes that the classifier recognizes.

Furthermore, after estimating s_j, the power of automatic contextual information extraction aids us in estimating the scene or room in an indoor environment, since scene estimation can be seen as a co-occurrence probability of the scene co-occurring with the objects recognized. This can be done by modifying the above formula and adding one more term *scene* in the query set, namely:

$$P(s_j, scene|c_1, c_2, ..., c_N) = \frac{1}{Z} \prod_{j,k} \psi(s_j, s_k) \prod_{j,i} \phi(s_j, c_i) \prod_{j,scene} \psi(s_j, scene) \quad (3)$$

where $j, k \in M$ and $i \in N$. In conclusion, by using contextual information the model becomes flexible enough so as not only to enhance the known classes' recognition, but also obtain a belief about novel observed objects and locations.

IV. RESULTS

In order to evaluate the performance of the proposed system, we carried out 2 different experiments. The first one regards the 3D classifier's recognition accuracy, whereas the second tests the quality of contextual information that we automatically extract from Flickr.

In the first experiment, we used 4 different classes of objects (chairs, tables, cupboards and sofas), as well as Internet-derived 3D CAD models from archive3d of the same classes. Every view of the self-recorded objects is considered as different object. We tested and trained upon different datasets. The confusion matrix in Table I shows the recognition's accuracy for each combination of training and testing.

We observed that training with only one type of data (kinect data only or web 3D models only) gives unstable results, especially when testing on other type of data. The best recognition outcomes arise after combining real-world data with web 3D models. This fact encourages us for complementing real-world datasets with online models in order to create multi-varied training sets and end up in more robust recognition.

Another interesting question that we wanted to answer is if we have a small self-recorded dataset, how much the augmentation of the 3D models would help in the recognition phase. For that reason we tested on our kinect database and then started to augment our training set with 3D models. In Figure 4 two different curves are showed:

- The red curve shows the recognition accuracy using only 3D models as training and augmenting them. As expected the recognition accuracy rises and gets balanced around 70%.
- We keep a fixed number of kinect data as training set and we augment with 3D models. The system is not affected too much by that augmentation and shows stability as expected from the confusion matrix in Table I.

TABLE I
CONFUSION MATRIX USING DIFFERENT DATASETS AS TRAIN AND TEST SETS.

Train \ Test	kinect	web 3d	kinect + web 3d
kinect	0.93	0.59	0.69
web 3d	0.68	0.90	0.83
kinect + web 3d	0.88	0.90	0.90

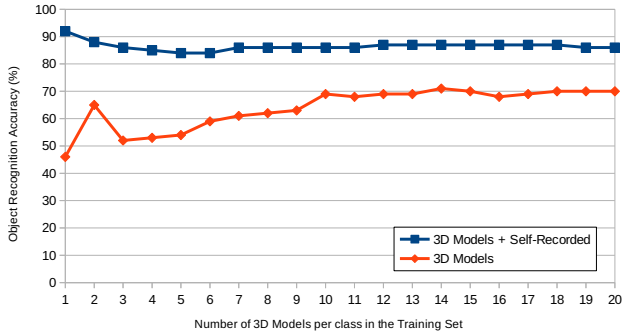

Fig. 4. Comparison of the object recognition accuracy between the classifier trained with self-recorded data and 3D CAD models and the one trained only with 3D models. Testing on the self-recorded data.

The second experiment tests the reliability of the co-occurrences probabilities computed from the web. In order to learn the co-occurrence probabilites, a FLICKR-derived database was used as detailed in Section III-C. To test the quality of the learning, the co-occurrence probabilites have been also computed from the Berkeley's Kinect-based 3D object dataset [37], which contains daily scenes, house rooms, office, etc. The FLICKR estimation is based on 351772 pictures containing 502 different terms (object classes) and the Berkeley's database is composed by 849 scenes with 83 different terms.

In Figure 5 a comparison between both co-occurrence probabilities distributions is shown for 6 different terms. This comparison shows that the co-occurrence probabilities distributions are similar for both databases. In spite of the very different purposes of the databases, the fact that both distributions are similar is remarkable. The objects of the Berkeley's database were manually tagged with scientific purposes. Whilst, FLICKR-derived database tagging is not strict since the pictures are tagged by the users who upload them, without following a common criteria in this process. The main conclusion of this comparison is that FLICKR can be used as a reliable source of context information, regardless

the automatic nature of our terms extraction.

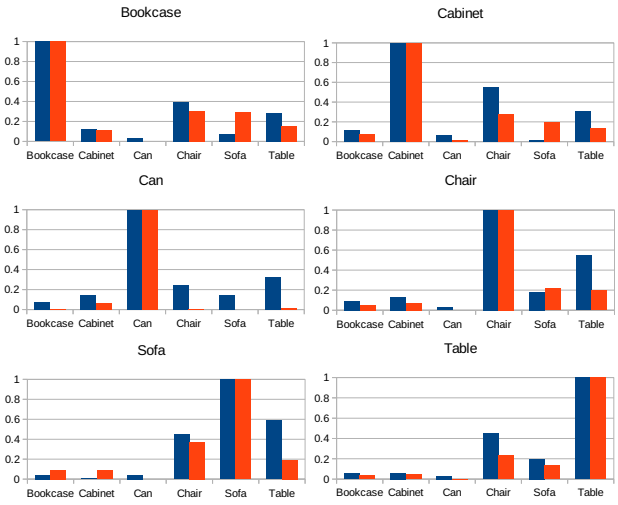

Fig. 5. Distributions of objects' co-occurrences probabilities in Flickr (orange) data and Kinectdata.com's (blue) database.

V. DISCUSSION AND FUTURE WORK

Traditional object recognition algorithms utilize objects' geometrical features in order to detect and categorize into different classes. However, several issues arise with these approaches, that we attempt to solve by introducing two novelties.

Firstly, there is a need for a labelled dataset of adequate size. The process of selecting and annotating big datasets is time-consuming and also hides always the possibility of limiting the recognition to specific environments and restricted conditions. We attempted to solve the size and annotation problem of the datasets by self-recording several objects and also downloading web 3D models from an online source. Afterwards, the features extraction phase followed for each object based on [4]. A nearest neighbour approach and *euclidean distance* was used for the final classification.

Secondly, in recent years, the object recognition community has realised that judging only from shape or color characteristics is not enough for robust results in real world situations. Occlusions and other environmental distortions affect the sensory perception of the robot making different objects look similar and hence any recognition task impossible. For that reason, we utilized context retrieved from social media websites and modelled them using markov random fields. The resulting classifier was probabilistic and has shown higher robustness and flexibility in relation to our initial 3D model based.

Regarding the integration of our system in a moving robot, we plan to incorporate one more active perception and learning step, especially on occasions where the robot meets unknown objects. That step is going to close the learning loop by re-training and extending the classifier. However, in order to attain smooth collaboration with humans, the support by a voice recognition software is essential. We further aim to test the resulting system against Internet-sized databases in order to provide an insight about its transferability.

Another interesting research direction for a robot that understands its environment, is to find optimal paths depending on the task that it wants to accomplish. For that reason, we plan to incorporate advanced navigation algorithms as well as information by other robots and media around for additional context. This is the last and necessary step before actually carrying out tasks in the real-world.

VI. CONCLUSION

For robots to be able to fluidly collaborate with and keep company to humans in indoor spaces, they need to be able to perceive and understand such environments, including furniture and rooms. Towards that goal, we presented a system for indoor furniture and room recognition for robots, which has two significant novelties: it utilizes internet-derived as well as self-captured models for training, and also uses object- and room-context information mined through the internet, in order to bootstrap and enhance its performance.

We have utilized silhouette-based features of Daras et al. [4] and a nearest neighbour approach for object recognition, and finally markov random fields for modelling object-object and object-scene context. Furthermore, in our approach we also used Internet-derived 3D models from archive3d and internet-mined co-occurrence probabilities through Flickr. Thus, the system also acts as an example of how autonomous robot entities can benefit from utilizing online information and services. Many interesting subproblems, including the peculiarities of utilizing such online sources, were discussed, followed by a real-world empirical evaluation of the system, which showed highly promising results.

ACKNOWLEDGEMENT

Parts of this research work have been produced using the EGI and HellasGrid infrastructures.

REFERENCES

[1] S. Golder and B. A: Huberman, "Usage Patterns of Collaborative Tagging Systems", *Journal of Information Science*, vol. 32, issue2, pp. 198-208, 2006.

[2] J. Mason and B. Marthi, "An Object-Based Semantic World Model for Long-Term Change Detection and Semantic Querying', *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Algarve, Portugal, To appear: 2012.

[3] H. Koppula, A. Anand, T. Joachims and A. Saxena, "Semantic Labeling of 3D Point Clouds for Indoor Scenes", *Neural Information Processing Systems (NIPS)*, Granada, Spain, 2011.

[4] P. Daras and A. Axenopoulos, "A 3D Shape Retrieval Framework Supporting Multimodal Queries", *International Journal of Computer Vision*, vol. 89, issue 2-3, pp 229-247, 2010.

[5] R. B. Rusu, G. Bradski, R. Thibaux, J. Hsu, "Fast 3D Recognition and Pose Using the Viewpoint Feature Histogram", *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, 2010.

[6] K. Lai and D. Fox, "3D Laser Scan Classification Using Web Data and Domain Adaptation", *Proceedings of Robotics: Science and Systems*, Washington, USA, 2009.

[7] H. Daumé, "Frustratingly easy domain adaptation", *In Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, 2007.

[8] D. Vranic, D. Saupe and J. Ritcher, "Tools for 3D-object Retrieval: Karhunen-loeve Transform and Spherical Harmonics", *Proceedings of the IEEE 4th Workshop on Multimedia Signal Processing*, pp. 293-329, 2001.

[9] J. Pu and K. Ramani, "An Approach to Drawing-like View Generation from 3D Models", *Proceedings of IDETC/CIE*, California, USA, 2005.

[10] J. Hays and A. Efros, "Scene Completion Using Millions of Photographs", *ACM Transactions on Graphics (SIGGRAPH)*, vol. 26, issue 3, 2007.

[11] J. Hays and A. Efros, 'IM2GPS: Estimating Geographic Information From a Single Image", *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, USA, 2008.

[12] J. Tighe and S. Lazebnik, "Superparsing: Scalable Nonparametric Image Parsing With Superpixels", *European Conference on Computer Vision*, Crete, Greece, 2010.

[13] A. Torralba, R. Fergus, and W. Freeman, "80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 30, issue 11, pp. 1958-1970 , 2008.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database", *European Conference on Computer Vision*, Princeton, USA, 2009.

[15] Google 3D Warehouse. http://sketchup.google.com/3dwarehouse (accessed 7 August 2012).

[16] S. Yingze Bao, M. Sun, and S. Savarese, "Toward Coherent Object Detection and Scene Layout Understanding", *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, 2010.

[17] D. Hoiem, A. Efros, and M. Hebert, "Recovering Surface Layout From an Image", *International Journal of Computer Vision*, vol. 75, issue 1, 2007.

[18] A. Saxena, M. Sun, and A. Ng, "Make3D: Learning 3D scene structure from a single still image", *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 31, issue 15, pp. 824-840 , 2009.

[19] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. Freeman, "SIFT Flow: Sense Correspondence Across Different Scenes", *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, USA, 2008.

[20] A. Oliva and A. Torralba, "Building the Gist of a Scene: the Role of Global Image Features in Recognition", *Progress in Brain Research*, 2006.

[21] M. Fisher and P. Hanrahan, "Context-based Search for 3D Models", *ACM Transactions on Graphics (SIGGRAPH)*, vol. 29, issue 6, 2010.

[22] M. Fisher, M. Savva, and P. Hanrahan, "Characterizing Structural Relationships in Scenes Using Graph Kernels", *ACM Transactions on Graphics (SIGGRAPH)*, vol. 30, issue 4, 2011.

[23] H. Grabner, J. Gall, and L. Van Gool, "What Makes a Chair a Chair?", *IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, USA, 2011.

[24] A. Nchter and J. Hertzberg , "Towards Semantic Maps for Mobile Robots", *Robotics and Autonomous Systems*, vol. 56, issue 11, pp. 915926, 2008.

[25] X. Xiong and D. Huber, "Using Context to Create Semantic 3D Models of Indoor Environments", *British Machine Vision Conference*, Guilford, UK, 2012.

[26] A. Oliva and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope", *International Journal of Computer Vision*, vol. 42, issue 3, pp.145-175, 2001.

[27] A. Oliva and A. Torralba, "The Role of Context in Object Recognition", *Trends in Cognitive Sciences*, vol. 11, issue 12, pp. 520-527, 2007.

[28] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert, "An Empirical Study of Context in Object Detection", *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, USA, 2009.

[29] 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 Proceedings of the 15th international Conference on Multimedia*, Augsburg, Germany, 2007.

[30] T. Quack, B. Leibe, L. Van Gool, "World-scale Mining of Objects and Events from Community Photo Collections", *In Proceedings of the 2008 international Conference on Content-Based Image and Video Retrieval*, Niagara Falls, Canada, 2008.

[31] D. Crandall, L. Backstrom, D. Huttenlocher, J. Kleinberg, "Mapping the Worlds Photos", *In Proceedings of the World Wide Web Conference*, Madrid, Spain, 2009.

[32] L. Kennedy and M. Naaman, "Less Talk, More Rock: Automated Organization of Community-Contributed Collections of Concert Videos", *In Proceedings of the World Wide Web Conference*, Madrid, Spain, 2009.

[33] X. Olivares, M. Ciaramita, R. van Zwol, "Boosting image retrieval through aggregating search results based on visual annotations", *In Proceeding of the 16th ACM international conference on Multimedia*, Vancouver, Canada, 2008.

[34] S. Nikolopoulos, E. Chatzilari, E. Giannakidou, A. Papadopoulos, I. Kompatsiaris, A. Vakali, "Leveraging Massive User Contributions for Knowledge Extraction", *Next Generation Data Technologies for Collective Computational Intelligence*, eds. Nik Bessis, Fatos Xhafa, Springer Verlag, Series: Studies in Computational Intelligence, 2011.

[35] E. Giannakidou, I. Kompatsiaris, A. Vakali, "SEMSOC: Semantics Mining on Multimedia Social Data Sources", *In Proceedings of the 2nd IEEE International Conference on Semantic Computing*, Santa Clara, USA, 2008.

[36] T. Kollar and N. Roy, "Utilizing Object-Object and Object-Scene Context when Planning to Find Things", *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Kobe, Japan, 2009.

[37] A. Janoch, S. Karayev, Y. Jia, J.T. Barron, M. Fritz, K. Saenko and T. Darrell, "A Category-Level 3-D Object Dataset: Putting the Kinect to Work", *International Conference on Computer Vision*, Barcelona, Spain, 2011.