IDIAP RESEARCH REPORT

TOWARDS SEMI-SUPERVISED LEARNING OF SEMANTIC SPATIAL CONCEPTS

Jesus Martinez-Gomez Barbara Caputo

Idiap-RR-03-2011

FEBRUARY 2011

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Towards Semi-Supervised Learning of Semantic Spatial Concepts

Jesus Martinez-Gomez and Barbara Caputo

Abstract— **The ability of building robust semantic space representations of environments is crucial for the development of truly autonomous robots. This task, inherently connected with cognition, is traditionally achieved by training the robot with a supervised learning phase. We argue that the design of robust and autonomous systems would greatly benefit from adopting a semi-supervised online learning approach. Indeed, the support of open-ended, lifelong learning is fundamental in order to cope with the dazzling variability of the real world, and online learning provides precisely this kind of ability. Here we focus on the robot place recognition problem, and we present an online place classification algorithm that is able to detect gap in its own knowledge based on a confidence measure. For every incoming new image frame, the method is able to decide if (a) it is a known room with a familiar appearance, (b) it is a known room with a challenging appearance, or (c) it is a new, unknown room. Experiments on a subset of the challenging COLD database show the promise of our approach.**

I. INTRODUCTION

Who wouldn't want a robot at home to make the daily chores? It could bring you a beer from the fridge, do the laundry, iron the shirts, collect things from the floor before cleaning, etc. A major requirement for having robots at home is that their representation of space, objects, and more generally concepts must at least partially overlap with our own. A vast literature in cognitive psychology (see [11] and reference therein) shows clearly that humans explain and categorize perceived multi-sensory patterns using semantic representations, of which language represents the synthesis. To fix ideas, let us focus here only on the semantic representation of space. We refer to rooms, and talk about them, in terms of their visual appearance (the corridor), the activities we usually perform in them (the fitness room) and the objects they contain (the bedroom). If we want to share our daily environment with robots, we need to share with them our own representation and understanding of it.

How do we make a robot learn the typical semantic space representation of humans? Robots have perceptual channels and cognitive abilities very different from our own. For instance, the typical service robot will use laser range scanners and an omnidirectional camera to collect data about an indoor place like an office environment. If programmed to learn the environment autonomously, i.e., in an unsupervised manner, the robot's interpretation of the data will result in a space representation very different from that of humans. Therefore, to make a robot have our own semantic representation of space, it is necessary to have a learning phase supervised by the user.

But how long should this supervised learning phase be? The current mainstream approaches (see Section II for a brief review of the relevant literature) assume a training phase, well separated from the actual working of the robot, where the human labels the data. Training usually stops when it is achieved a pre-defined threshold level of performance on a validation set of data, or when the user decides it. From that moment on, the robot is on its own. We argue that this approach is doomed to fail: rooms change around us continuously over time as furniture is added, replaced or relocated. It is impossible to predict how a user is going to redecorate its living room in the future, and therefore it is impossible to train the robot beforehand on such data.

Our vision is that the supervised learning mode should always be accessible to the robot, and it should be triggered by its ability to explain the incoming data. The transition from fully supervised to unsupervised should be smooth, robot driven, and competence-based. In other words, our vision is that semi-supervised online learning should become the mainstream approach for enabling robots to learn semantic concepts.

To move towards this goal, here we present an algorithm able to learn semantic spatial concepts in an open ended fashion, i.e. continuously updating its internal model with a bounded memory growth. The robot switches from a fully autonomous, unsupervised learning phase to a supervised one (where assistance by a human teacher might be required) on the basis of its capability to interpret the data with a high degree of confidence. The capability to detect hard-to-explain incoming data is done at the classifier level, frame by frame, as well as at a higher level, by exploiting the temporal continuity of the image sequences. This permits to distinguish between challenging instances of a known spatial concept (a view of the known class kitchen where it is perceived for the first time a new piece of furniture) and a new concept (a room never seen before).

Concretely, our algorithm consists of two components: the first is an online learning algorithm with performance comparable to that of the batch method and a bounded memory growth; the second is a mechanism for assigning labels to incoming data, detecting challenging frames imaging known concepts and ultimately recognizing when being in a whole new room. We take an discriminative approach and we build on previous work on online learning [18], [25] and confidence-based place classification [19]. Experiments

This work was supported by the Spanish Junta de Comunidades de Castilla-La Mancha under PCI08-0048-8577 and PBI-0210-7127 projects (J. M.-G.) and by the SS2Rob project (B.C.). The support is gratefully acknowledged.

J. Martinez is with the I3A Research Institute, Campus Universitario s/n, 02071, Albacete, Spain jesus_martinez@dsi.uclm.es

B. Caputo is with the Idiap Research Institute, Rue Marconi 19, 1920 Martigny, Switzerland bcaputo@idiap.ch

on a subset of the challenging COLD database [20] show promising results.

The rest of the paper is organized as follows: after a brief review of the related literature, we describe the two components of our approach: the online learning algorithm III and the detection of confidence/ignorance IV. Section V describes our experimental setup, while section VI reports our experimental findings. We conclude with an overall discussion and possible future avenues for research.

II. RELATED WORKS

The ability to learn and interpret complex sensory information based on previous experience, inherently connected with cognition, has been recognized as crucial and vastly researched [23], [21], [16]. In most cases, the recognition systems used are trained offline, i.e., they are based on batch learning algorithms. However, in the real dynamic world, learning cannot be a single act. It is simply not possible to create a static model which could explain all the variability observed over time. Continuous information acquisition and exchange, coupled with an ongoing learning process, is necessary to provide a cognitive system with a valid world representation.

In the last few years, the need for solutions to such problems as the robustness to long-term dynamic variations, or the transfer of knowledge, is more and more acknowledged. In [21], the authors tried to deal with long-term visual variations in indoor environments by combining information acquired using two sensors of different characteristics. In [26], the problem of invariance to seasonal changes in appearance of an outdoor environment is addressed. Clearly, adaptability is a desirable property of a recognition system. At the same time, Thrun and Mitchell [24], [15] studied the issue of exchanging knowledge related to different tasks in the context of artificial neural networks and argued for the importance of knowledge-transfer schemes for lifelong robot learning. Several attempts to solve the problem have also been made from the perspective of Reinforcement Learning, including the case of transferring learned skills between different RL agents [14], [10].

III. STEP 1: MEMORY CONTROLLED ONLINE LEARNING AND RECOGNITION OF VISUAL PLACES

This section describes the first component of our overall approach, namely an online learning algorithm with a bounded memory growth and an accuracy comparable to the classic, off-line method. We take a discriminative approach, and derive an approximate version of the Online Independent-SVM. As opposed to the original algorithm, our approach does not require to store all incoming data but it allows to discard most of them in a principled manner. This leads to a bounded memory growth, where the upper bound is set by the user and the lower bound by theoretical constraints. In the rest of this section we first review basic concepts on SVM (section III-A), then we summarize the OI-SVM algorithm (section III-B). Our Memory Controlled OI-SVM is described in section III-C.

A. SUPPORT VECTOR MACHINES

Due to space limitations, this is a very quick account of SVMs — the interested reader is referred to [3] for a tutorial, and to [6] for a comprehensive introduction to the subject. Assume $\{\mathbf{x}_i, y_i\}_{i=1}^{l}$, with $\mathbf{x}_i \in \mathbb{R}^m$ and $y_i \in \{-1, 1\}$, is a set of samples and labels drawn from an unknown probability distribution; we want to find a function $f(\mathbf{x})$ such that $sign(f(\mathbf{x}))$ best determines the category of any future sample $\mathbf{x}$. In the most general setting,

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{1}$$

where $b \in \mathbb{R}$ and $K(\mathbf{x}_1, \mathbf{x}_2) = \Phi(\mathbf{x}_1) \cdot \Phi(\mathbf{x}_2)$, the kernel function, evaluates inner products between images of the samples through a non-linear mapping Φ. The α_is are Lagrangian coefficients obtained by solving (the dual Lagrangian form of) the problem

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \tfrac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{l} \xi_i^p \\ \text{subject to} \quad & y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{2}$$

where $\mathbf{w}$ defines a separating hyperplane in the feature space, i.e., the space where Φ lives, whereas $\xi_i \in \mathbb{R}$ are slack variables, $C \in \mathbb{R}^+$ is an error penalty coefficient and p is usually 1 or 2. In practice, most of the α_i are found to be zero after training; the vectors with an associated α_i different from zero are called support vectors. Notice that, from (1), the testing time of a new point is proportional to the number of SVs, hence reducing the number of SVs implies reducing the testing time.

B. ONLINE INDEPENDENT SUPPORT VECTOR MACHINES

Let the *kernel matrix* K be defined such that $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, with $i, j = 1, \ldots, l$. The possibility to obtain a more compact representation of $f(\mathbf{x})$ follows from the fact that the solution to a SVM problem (that is, the α_is) is not unique if K does not have full rank [3], which is equivalent to some of the SVs being linearly dependent on some others in the feature space [8]. Orabona et al [18] applied this idea to the online learning framework. As it would be unfeasible a simplification of the solution each time a new sample is acquired, they suggested to use independent SVs only, that is to decouple the concept of "basis" vectors, used to build the classification function (1), from the samples used to evaluate the ξ_i in (2). If the selected basis vectors span the same subspace as the *whole sample set*, the solution found will be equivalent.

The OI-SVM algorithm adds incrementally a new incoming samples if it is linearly independent in the feature space from those already present in the basis itself. The solution

found is *the same* as in the classical SVM formulation; therefore, no approximation whatsoever is involved.

Denoting the indexes of the vectors in the current basis, after l training samples, by $\mathcal{B}$, and the new sample under judgment by $\mathbf{x}_{l+1}$, the algorithm can then be summed up as follows:

1) check whether $\mathbf{x}_{l+1}$ is linearly independent from the basis in the feature space; if it is, add it to $\mathcal{B}$; otherwise, leave $\mathcal{B}$ unchanged.
2) incrementally re-train the machine.

Hence the testing time for a new point will be $O(|\mathcal{B}|)$, as opposed to $O(l)$ in the standard approach; therefore, keeping $\mathcal{B}$ small will improve the testing time without losing any precision whatsoever. A major drawback of OI-SVM is that it requires to store in memory all the incoming training data in order to guarantee that the online solution is the same as in the classical SVM formulation.

C. MEMORY CONTROLLED ONLINE INDEPENDENT SUPPORT VECTOR MACHINES

The need to store all incoming data makes in practice unusable the OI-SVM algorithm for open-ended learning of semantic spatial concepts, especially for a mobile robot platform: while the dimension of the solution would remain constant over time, the overall memory requirement would grow linearly with the number of perceived frames, leading quickly to a memory explosion.

To overcome this problem, we propose to apply a forgetting strategy over the stored Training Samples (TSs), while preserving the stored Support Vectors in order to approximate reasonably well the original optimal solution. The idea of keeping under control the memory growth of online learning algorithms is not new: several authors tried in the past to address this problem, mainly by bounding a priori the memory requirements. The first algorithm to overcome the unlimited growth of the support set was proposed by Crammer et al. [5]. The algor ithm was then refined by Weston et al. [27]. The idea of the algorithm was to discard a vector of the solution, once the maximum dimension has been reached. The strategy was purely heuristic and no mistake bounds were given. A similar strategy has been used also in NORMA [9] and SILK [4]. The very first online algorithm to have a fixed memory "budget" and at the same time to have a relative mistake bound has been the F orgetron [7]. Within the context of semantic scene recognition, Ullah et al [25] proposed instead a random forgetting strategies, which should be more robust to possible unbalancing into the class-by class distribution of the TSs.

Here we take the approach proposed in [IROS09] and define the following random forgetting strategy:

1) we introduce a threshold value that corresponds to the allowed maximum number of stored Training Samples ($MaxTSs$);
2) whenever $TS > MaxTSs$, we randomly discard TSs until their value is again below threshold. This concretely means discarding old TSs, selected randomly, for each new incoming TS.
3) With this strategy, the memory requirements of the algorithm are always between the number of SVs of the testing solution and the number of SVs plus $MaxTSs$.

We will show experimentally that this approximation of the original OI-SVM algorithm does not affect the accuracy of the solution for a wide range of values of $MaxTSs$.

IV. STEP 2: DETECTION OF IGNORANCE

The second, key component of our method is the capability to autonomously assign labels to new, incoming images, without the need for human supervision. The core issue here is the ability to estimate the level of confidence of each potential decision: a frame classified as corridor should be used to update the internal representation for the class corridor only if the confidence of the decision is high enough. If this would not be the case, then there would be a very strong risk of adding wrongly labelled data to the model, with a consequent degradation of the overall performance over time.

At the same time, one could argue that the most challenging frames, for each known class, are the most important to be added as they are those bringing new valuable information. An obvious way to do so would be to store the challenging frames and then, periodically, asking for labels to a human supervisor. Our solution here is instead to exploit the temporal continuity between frames and the intrinsic constraints of the problem: once a robot has traversed a door, all frames perceived until crossing another door must belong to the same semantic spatial concept. This same line of reasoning gives us a useful tool to determine if the robot has entered a new, unknown room.

Section IV-A describes how we estimate the level of confidence of the classifier and how to exploit temporal continuity to label challenging frames. Section IV-B illustrates how these two ingredients can be also used to identify new semantic spatial concepts.

A. DETECTING CHALLENGING FRAMES

To use incoming data to update the internal models, the algorithm needs to assign reliably class labels to each new frame. This in turns means that it should be able to detect frames that cannot be properly classified, i.e. that cannot be classified with a high confidence level. The problem therefore becomes that of defining effective confidence measures for evaluating the reliability of the label assignment process. As we use a multiclass SVM with one versus all strategy, a natural measure of confidence is the decision margin for each class. These margins will be positive when a frame should be classified using one of the known classes (hard acceptance), negatives otherwise (hard rejection). Figure 1, right, shows an example of the margins output for the class corridor in case of a frame correctly classified with high confidence.

On the basis of the output margins $M_{ni=1}^{iC}$, with C= number of classes, for each frame n, we define the two following conditions for detecting challenging frames:

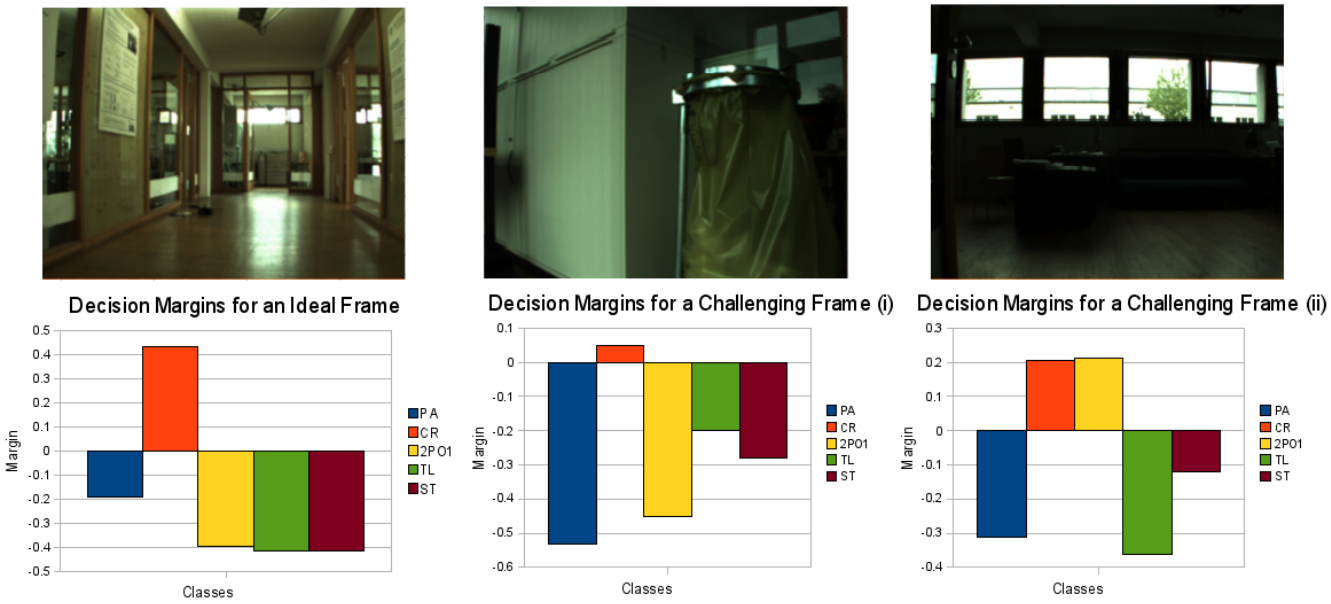


Fig. 1. Decision margins obtained for an ideal frame and two types of challenging frames.

1) $M_n^i < M_{max}{}_{i=1}^{C}$: for each of the possible classes C none obtains a high level of confidence;
2) $|M_n^i - M_n^j| < \Delta_{i=1}^{C}$: there are at least two classes with high level of confidence, but their absolute difference is too small to allow for a confident decision.

Figure 1, centre, shows an example of a frame classified as challenging because of a low level of confidence (condition (1)); Figure 1, left, shows instead an example of challenging frame where there are two high and very close levels of confidence (condition (2)).

To improve stability, we normalize the margins by dividing all values by the maximum positive (if the margin is positive) or the lowest negative value (otherwise). The obtained set of margins will be within $(-1.0, +1.0)$. The threshold values (M_{max}, Δ) are of course crucial for the success of the method. In Section VI we show experiments exploring the robustness of the method to these two parameters.

Once a frame has been identified as challenging, we use the classification results obtained for the last n frames to solve the ambiguity: if all the last n frames have been assigned to the class C_i, then we can conclude that all frames come from the same class C_i, and the label will be assigned accordingly. This could be further integrated with a door detection algorithm to avoid false label assignments. We do not purse here this idea, although we plan to do it in the future.

B. DETECTING NEW ROOMS

A special type of challenging frames are those corresponding to new rooms. When robots enter into a room not seen during training, we would expect that most of the margin values for all known classes should be negative, or anyway with low positive confidence. Furthermore, we would expect that by looking at n consecutive frames one would not be able to detect a dominant class label. In such a case, all incoming frames will be detected as challenging.

When such a situation (all test frames classified as challenging frames) continues for a large number of frames, we consider that the robot has entered a new room, not seen during training. In such a case, the robot can only ask for labels to a human supervisor. Here the critical parameter is of course the minimum number of challenging frames to be detected continuously: this point has been investigated experimentally in section VI.

V. EXPERIMENTAL SETUP

In this section we describe the experimental setup used to validate our approach. Section V-A describes the data used, and section V-B the feature descriptors. The description of each experiment with the corresponding result is given in section VI.

A. THE DATABASE

For all our experiments we used a subset of the COLD database [20]. It contains three separate sub-datasets, acquired at three different indoor labs, located in three different European cities: the Visual Cognitive Systems Laboratory at the University of Ljubljana, Slovenia; the Autonomous Intelligent System Laboratory at the University of Freiburg, Germany; and the Language Technology Laboratory at the German Research Center for Artificial Intelligence in Saarbrücken, Germany. For each lab, image sequences of several rooms are provided, all acquired with the same camera settings.

Here we used the sub-dataset acquired in the Autonomous Intelligent System Laboratory at the University of Freiburg, Germany (COLD-Freiburg): it consists of three sets of sequences, both acquired under varying illumination conditions. Of these three sets, we chose the following two: In

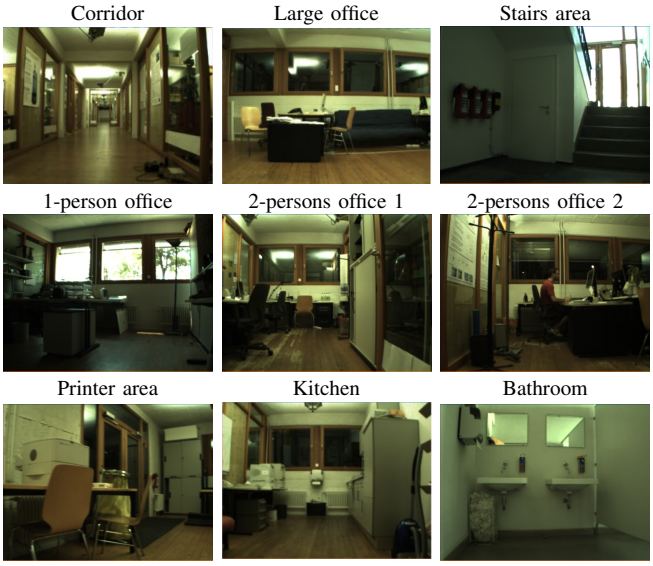


Fig. 2. Examples of images from the COLD-Freiburg database.

the first set, the robot travels across five rooms: corridor, 2-person office, printer area, bathroom and stairs area. In the second set, the robot travels across the rooms of the first set, plus other four rooms: a 1-person office, a printer area, a kitchen and a large office. Figure 2 shows some exemplar views from the second set of sequences. Each of the sequences described above were acquired under three different illumination conditions -sunny, cloudy and night. Three sequences were acquired, one after the other, for each weather condition, for a total of nine data sequences for each set.

B. THE FEATURES

As features, we chose a variety of global descriptors representing different features of the images. We opted for histogram-based global features, mostly in the spatial-pyramid scheme introduced in [12]. This representation scheme was chosen because it combines the structural and statistical approaches: it takes into account the spatial distribution of features over an image, while the local distribution is in turn estimated by mean of histograms; moreover it has proven to be more versatile and to achieve higher accuracies in our experiments.

The descriptors we have opted to extract belong to five different families: Pyramid Histogram of Orientated Gradients (PHOG) [2], Sift-based Pyramid Histogram Of visual Words (PHOW) [1], Pyramid histogram of Local Binary Patterns (PLBP) [17], Self-Similarity-based PHOW (SS-PHOW) [22], and Compose Receptive Field Histogram (CRFH) [13]. Among all these descriptors, CRFH is the only one which is not computed pyramidly. For the remaining families we have extracted an image descriptor for every value of $L = \{0, 1, 2, 3\}$, so that the total number of descriptors extracted per image is equal to 25 ($4 + 4$ PHOG, $4 + 4$ PHOW, 4 PLBP, 4 SS-PHOW, 1 CRFH). In order to select the best visual cues to be combined together we performed a pre-selection step, namely we run some preliminary experiments to decide which combination of features was more effective. This eventually made us settle on two descriptors, PHOG L0 and Oriented PHOG L2. Their exact settings are summarized in Table I. These two features are concatenated to generate a single feature that will be used as input for the classifier.

DESCRIPTOR	SETTINGS	L
PHOG_{180}	range= $[0, 180]$ and $K = 20$	$\{0\}$
PHOG_{360}	range= $[0, 360]$ and $K = 40$	$\{2\}$

TABLE I
SETTINGS OF THE IMAGE DESCRIPTORS

VI. RESULTS

This section presents an experimental evaluation of our approach. We first test the performance of MC-OI-SVM compared to that of the original method (section VI-A). Then we analyze the impact of detecting challenging frames on the overall performance, especially in terms of false positives (section VI-B). Lastly, we investigate the capability to detect unknown rooms (section VI-C).

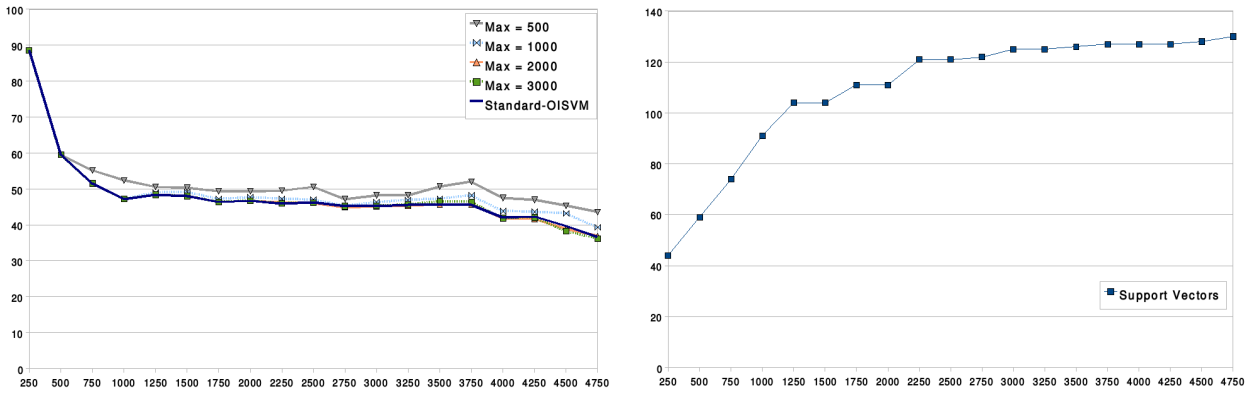
For all the experiments, we used the COLD-Freiburg database and the visual features described in the previous section. Training always consisted of three sequences, acquired one after the other, with the same illumination conditions. Testing consisted of one sequence, taken from those not used for training. For the SVM, we used the χ^2 kernel, with $C = 1$, $\gamma = 1$ and $\eta = 0.25$.

A. Experiment 1: Memory-Controlled OISVM

To compare the performance of MC-OI-SVM with that of OI-SVM, we used only sequences from the A set of the COLD-Freiburg, i.e. the testing sequences did not contained rooms not seen during training. Because of the different illumination conditions (cloudy, sunny and night) and of the number of acquired sequences for each lighting setting, there are 9 different combinations of training and test data. We performed experiments on all of them, selecting for MC-OI-SVM four different values of $MaxTSs$: (500,1000,2000,3000).

In order to show how this threshold affects the memory requirements of the system, Fig 3 shows the number of training samples stored by the algorithm, for all different values used for $MaxTSs$.

Fig. 4 shows the obtained results, averaged over the 9 runs. Figure 4, left, shows the error rate for the two algorithms, for the different values of $MaxTSs$. We see that, for $MaxTSs \geq 1000$, the performance of the two algorithms is essentially the same. Figure 4, right, shows instead the number of stored vectors for the testing solution, for all methods. From this figure, and from Fig 3, we can make two remarks: (1) with MC-OI-SVM it is possible to obtain an impressive reduction on the memory requirements of the original method with a very negligible decrease in accuracy; (2) when the $MaxTSs$ value is too close to the number



Error Rate Number of Support Vectors

Fig. 4. Results obtained for experiment training with images acquired under cloudy conditions

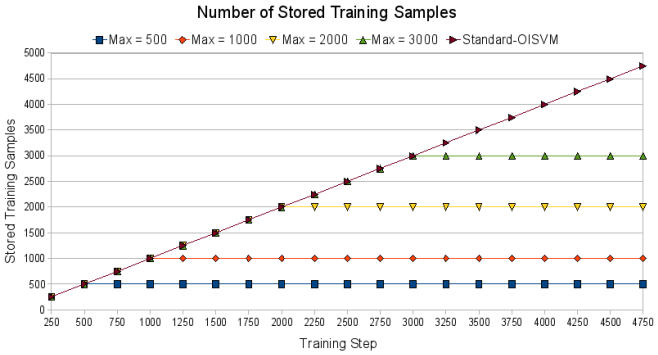


Fig. 3. Number of stored training samples for different $MaxTSs$ values

of SV of the testing solution, the approximation affects the optimality of the solution, and the error rate starts to increase.

These results confirm the effectiveness of our online approach; in the following, we will use MC-OI-SVM only, with $MaxTSs = 2000$.

B. Experiment 2: Detecting Challenging Frames

We now turn to analyze the capability of our approach to detect challenging frames. Training and test data were chosen as described in the previous section, but here we decided to show results relative to only one specific run (training on night, testing on cloudy) for the sake of clarity. Results obtained on the other 8 runs are similar and omitted here for space reasons.

Figure 5, left, shows the accuracy obtained for different values of the confidence threshold, with $\Delta = 0$, in terms of true positives, false positives and frames not classified, corresponding to the challenging frames. We see that, for threshold values between 0.05 and 0.1, almost all the frames detected as challenging are false positives of the original approach. When the threshold values increases, also the number of true positives start decreasing.

Figure 5, right, shows the effect of using the temporal continuity ('history factor'): for a threshold value of 0.2, we see that not only the number of true positives increases considerably, but that up to 25 consecutive frames the number

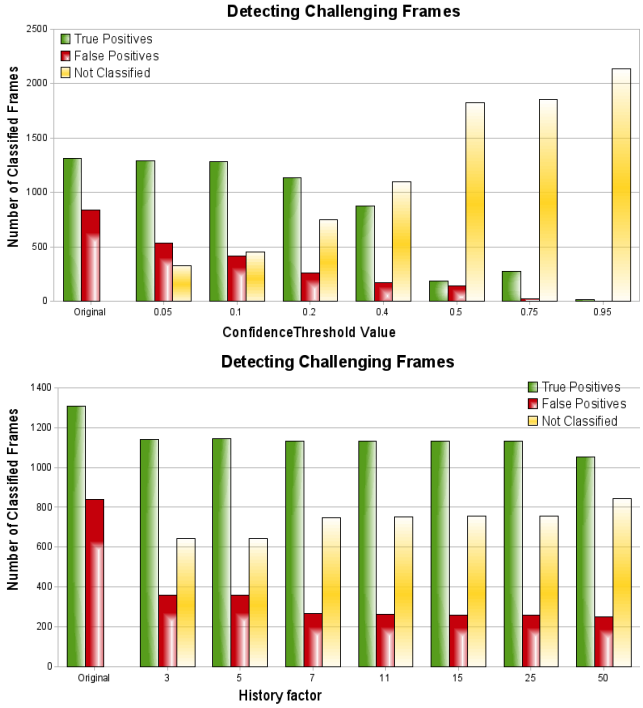


Fig. 5. Challenging Frames detection with different threshold values

of detected challenging frames increases almost exclusively at the expenses of the false positives.

We also tested different values of Δ, but we did not observe any improvement in the performance, given that all the experiments we run on data collected in the same indoor environment, we believe that this result cannot be considered conclusive for an evaluation of its usefulness.

C. Experiment 3: Detecting New Rooms

As a last, final set of experiments, we tested the capability of our system to distinguish between challenging frames of a known room, and a new room never seen during training. To this end, we used as testing sequences the B sequence of the COLD-Freiburg database, considering as new rooms

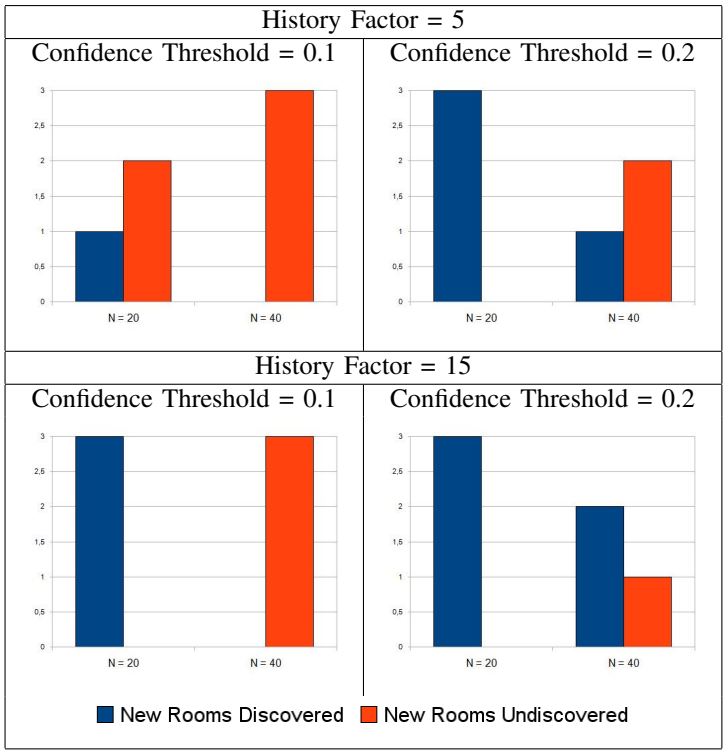


Fig. 6. Number of unknown rooms detected for different combinations of parameters

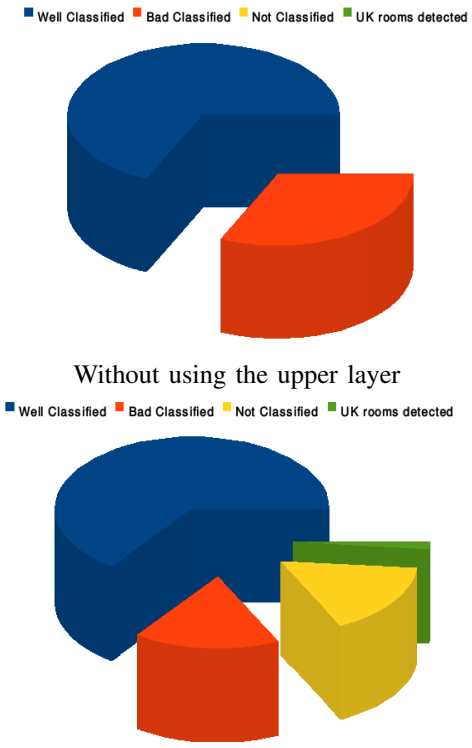


Fig. 7. Processed Output using the upper layer

three of the extra four images (we considered the two printer areas as the same categorical room). Figure 6 shows the results obtained, for two possible threshold values (0.2 and 0.4) and two different values of the history factor (5 top, 15 bottom). When using a history factor of 5 and a confidence level of 0.1, the system recognizes only one of the three rooms as unknown, after 20 consecutive challenging frames (Figure 6, top left). When, one considers instead at least 40 consecutive challenging frames, the system is not able to recognize any unknown room. A similar behaviour is observed with a history factor of 15: when passing from 20 to 40 consecutive challenging frames, the system passes from detecting all the unknown rooms to none (Figure 6, bottom left). The same behavior, slightly less accentuated, can be observed with a threshold value of 0.2 (Figure 6, right). This makes us conclude that it is better to ask for a smaller number of consecutive challenging frames, and a slightly high confidence threshold.

Finally, Fig. 7 presents graphically the results obtained using or not the upper layer –the advantage is quite evident.

VII. CONCLUSIONS AND FUTURE WORK

In this paper we presented an algorithm for online learning of semantic spatial concepts with a bounded memory growth, able to measure its own level of confidence when classifying incoming frames, and therefore able to decide when to ask for human annotation and when to trust its own decisions. Experiments on a subset of the challenging COLD database [20] show that our approach is able to minimize the false positives when classifying known frames, and it is able to detect new rooms, not seen during training.

Besides a more extensive experimental evaluation, this work can be continued in many ways. With respect to the confidence estimate, here we used the output margin of the SVM-based classifiers, but more elegant and sophisticated options should be explored here. We plan to integrate the high level layer of the approach with a door detector, so to increase the robustness of the process. Lastly, here we applied the method to only visual features, but this framework should work, and benefit from, multi-modal data such as laser range features. Future work will proceed in these directions.

REFERENCES

[1] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *International Conference on Computer Vision*. Citeseer, 2007, pp. 1–8.

[2] ——, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007, p. 408.

[3] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Knowledge Discovery and Data Mining*, vol. 2, no. 2, 1998.

[4] L. Cheng, S. V. N. Vishwanathan, D. Schuurmans, S. Wang, and T. Caelli, "Implicit online learning with kernels," in *Advances in Neural Information Processing Systems 19*, 2007.

[5] K. Crammer, J. Kandola, and Y. Singer, "Online classification on a budget," in *Advances in Neural Information Processing Systems 16*, 2003.

[6] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods)*. CUP, 2000.

[7] O. Dekel, S. Shalev-Shwartz, and Y. Singer, "The Forgetron: A kernel-based perceptron on a budget," *SIAM Journal on Computing*, vol. 37, no. 5, pp. 1342–1372, 2007.

[8] T. Downs, K. E. Gates, and A. Masters, "Exact simplification of support vectors solutions," *Journal of Machine Learning Research*, vol. 2, pp. 293–297, 2001.

[9] J. Kivinen, A. Smola, and R. Williamson, "Online learning with kernels," *IEEE Trans. on Signal Processing*, vol. 52, no. 8, pp. 2165–2176, 2004.

[10] G. Konidaris and A. G. Barto, "Autonomous shaping: knowledge transfer in reinforcement learning." in *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[11] G. Lakoff, *Women, fire and dangerous things: what categories reveal about the mind*. The University of Chicago Press, 1990.

[12] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006.

[13] O. Linde and T. Lindeberg, "Object recognition using composed receptive field histograms of higher dimensionality," in *Proc. ICPR*. Citeseer, 2004.

[14] J. Malak, R.J. and P. K. Khosla, "A framework for the adaptive transfer of robot skill knowledge using reinforcement learning agents," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'01)*, 2001.

[15] T. Mitchell, "The discipline of machine learning," CMU, Tech. Rep. CMU-ML-06-108, 2006.

[16] A. C. Murillo, J. Kosecka, J. J. Guerrero, and C. Sagues, "Visual door detection integrating appearance and shape cues," *Robotics and Autonomous Systems*, vol. 56(6), pp. pp. 512–521, June 2008.

[17] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Gray scale and rotation invariant texture classification with local binary patterns," *Computer Vision-ECCV 2000*, pp. 404–420, 2000.

[18] F. Orabona, C. Castellini, B. Caputo, J. Luo, and G. Sandini, "Indoor place recognition using online independent support vector machines," in *Proc. BMVC*, vol. 7. Citeseer.

[19] A. Pronobis and B. Caputo, "Confidence-based cue integration for visual place recognition," *Proc. IROS07*.

[20] ——, "COLD: COsy Localization Database," *The International Journal of Robotics Research (IJRR)*, vol. 28, no. 5, May 2009. [Online]. Available: http://www.csc.kth.se/ pronobis/research/pronobis09ijrr-cold/pronobis09ijrr-cold.pdf

[21] A. Pronobis, O. Martínez Mozos, and B. Caputo, "SVM-based discriminative accumulation scheme for place recognition," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'08)*, Pasadena, CA, USA, May 2008.

[22] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, 2007, pp. 1–8.

[23] C. Siagian and L. Itti, "Biologically-inspired robotics vision monte-carlo localization in the outdoor environment," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, San Diego, CA, USA, October 2007.

[24] S. Thrun and T. Mitchell, "Lifelong robot learning," *Robotics and Autonomous Systems 15*, 1995.

[25] M. Ullah, F. Orabona, B. Caputo, I. IRISA, and F. Rennes, "You live, you learn, you forget: Continuous learning of visual places with a forgetting mechanism," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009. IROS 2009*, 2009, pp. 3154–3161.

[26] C. Valgren and A. J. Lilienthal, "SIFT, SURF and seasons: Long-term outdoor localization using local features," in *Proceedings of the European Conference on Mobile Robots (ECMR'07)*, 2007.

[27] J. Weston, A. Bordes, and L. Bottou, "Online (and offline) on an even tighter budget," in *Proceedings of AISTATS 2005*, R. G. Cowell and Z. Ghahramani, Eds., 2005, pp. 413–420.