

Indoor Object Recognition through Human Interaction using Wavelet Features

QingHua Wang¹

Luís Seabra Lopes^{1,2}

¹ *Instituto de Engenharia Electrónica e Telemática de Aveiro*

² *Departamento de Electrónica e Telecomunicações
Universidade de Aveiro, 3810-193, Aveiro, Portugal
qhwang@ieeta.pt, lsl@det.ua.pt*

Abstract

In this paper a preliminary work towards grounded concept learning for a service robot through its vision and human interaction is presented. With a lifelong learning server (LLL), described in [8], the robot can incrementally learn to recognize instances of such concepts of indoor objects as “Person”, “Trash-can” and “Triangle sign” using simple intra-band statistical features extracted from the Haar wavelet transform of its vision information under the instruction of a human teacher. Experimental results show that these simple wavelet-based features can efficiently describe the characteristics of different objects in an office-like environment. Comparison with some other feature extraction methods is also given.

1 Introduction

Although many applications have been reported to address automatic object recognition under constrained conditions, such as controlled illuminations or backgrounds, it's still an open problem waiting for new efforts especially in unconstrained environments since it's usually crucial premise for real applications concerning symbol grounding [4] and image understanding. It also fits our project, CARL, which aims to contribute to the development of intelligent service robots. The results of this project include a prototype of an intelligent service robot, called Carl¹, which can execute such tasks as serving food in a reception or acting as a host in an

organization [7, 10]. As it can be found in these robot-human interactions, object recognition is the first step of concept learning. As a dimensionality reduction and feature transformation method, wavelet transform runs faster (in time complexity of $O(n)$) than such methods as Principle Component Analysis, Linear Discriminant Analysis, Factor Analysis or DCT / Blocked DCT. This is very crucial, especially in (near) real-time applications. In the context of object detection and recognition, the wavelet transform is believed to be good at localizing edges and other anomalies. In our work, wavelet transform not only serves for image representation, but also for dimension reduction. The simplest and fastest wavelet transforms, Haar wavelet transform, is used in the work reported below.

There are already several applications applying Haar wavelet into object detection and recognition. In [6], a “overcomplete dictionary” of Haar wavelet basis functions, used in characterizing the specific object classes, is learned from the class instances, based on an empirical statistical analysis. Then the class model is learned from these selected basis functions by support vector machines (SVMs). In [14], to detect vehicles on road, a 5 level Haar wavelet is applied on the detected vehicle candidate and all coefficients except those in the HH subband of the first level transform are directly used as input to SVMs and thus a decision is given.

There are also some approaches using some simple but informative features extracted from wavelet space. In [1], A feature extraction approach based on Daubechies wavelet was used to analyze the audio tracks accompanying videos. Five parameters were computed from the wavelet space, namely *centroid* and *bandwidth* of the whole wavelet space along with *energy*, *variance*

¹ Project “CARL: Communication, Action, Reasoning and Learning in Robotics”, FCT PRAXIS /12121/98

and *zero crossing rate* in each subband. In [2, 3], mean values and the corresponding variances were computed from face images to characterize different faces. Specifically, 3 mean values and 3 variances of the approximation image and 15 variances of details images form the feature vector. Then the Bhattacharyya distance was used to classify the feature vectors into person classes.

A major advantage of extracting and using a subset of features from wavelet space is to speed up learning processes. For the near real-time requirement in our project, we only use standard first order statistical features such as mean value, energy and variance of each wavelet subband. We call them “intra-band” information. Here we present an approach for indoor object recognition using this intra-band information. Unlike experiments carried out in [2, 3], we use unconstrained images of objects in a normal office. The experimental results show that these features work quite well in our application to recognize the instances of indoor objects. As we may notice, there is a semantic gap between the features extracted by the robot from the samples and concepts of the objects in those samples. Since these concepts are semantic terms used by humans to identify them, we hope the direct integration of the human user into the learning process of the robot can address this problem.

The rest of this paper is organized as follows. The proposed approach is described in Section 2. The Haar wavelet theory is briefly presented in Section 3. Feature extraction from wavelet space is provided in Section 4. Section 5 describes the experiments and obtained results. Comparison with our previous work is also presented in this section. Some discussion and future work is provided in Section 6 and in Section 7 conclusions are given.

2 System Overview

Generally speaking, robot learning should be seen as a lifelong process [8, 9]. The focus is to assign robots a kind of capability to self-development with rich sensory-motor skills and minimal initial knowledge, rather than to assign robots rich knowledge for task solving in advance. See [11, 15] for more details. After the construction phase, a lifelong learning process starts. The robots can learn new skills, explore environments themselves or under the guidance from human interactants.

On-line lifelong learning in robotics has seldom been described. Moreover, a few known systems demonstrating on-line learning, still are mostly limited to subsymbolic learning and/or don't address the grounding problem. In contrast, we are extending Carl's learning capabilities in order to support grounding of natural language concepts for human-robot interaction. Currently, the focus is on visually recognizing indoor objects, such as "person",

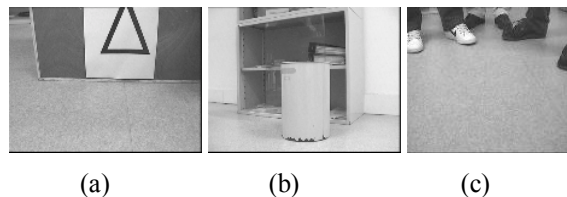


Figure 1. Sample images of “Person”, “Trash Can” and “Triangle Sign” respectively (re-sized her).

“trash can” or “triangle sign”. Figure 1 shows images containing instances of these concepts. Figure 2 provides the overview of our approach. The idea is that, when Carl meets an obstacle, he may ask, e.g.:

“Is this a person?”

Based on the obtained answer (“yes” or “no”) and the visual information, Carl may store a labeled example. A collection of labeled examples like this will enable it to induce the concept of “person”. The approach, therefore, consists of enabling Carl to manage incrementally and concurrently multiple learning problems through a learning server, called LLL [8]. We are currently using a 3-level Haar transform as the basis for obtaining an informative subset of the huge vision space that improves learning performance.

As far as communication with human instructors is concerned, Carl is able to enter a spoken language conversation with a human user in English. Speech recognition is currently based on NUANCE. The recognition grammar being used is able to accept over 12000 different sentences [7, 9].

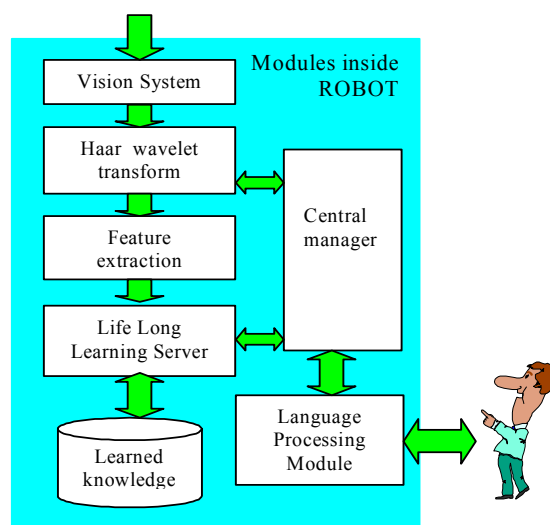


Figure 2. Approach Overview

3 Haar Wavelet

Although there is specialized literature describing wavelet transformation [5, 12, 13], we present here a brief description, especially of Haar wavelet transformation. The essence behind wavelets is to analyze arbitrary signals according to its scales in frequency domain. Thus it's a type of multiresolution analysis. Wavelets are functions defined over a finite interval. They are obtained from a single prototype wavelet called mother wavelet by dilation and translation at different positions and on different scales. So arbitrary signals can be represented as a linear combination of such wavelets, or basis functions.

We can formalize the notion of a multiresolution analysis as a nesting of the spanned subspaces:

$$V^0 \subset V^1 \subset V^2 \subset \dots \subset V^j \subset V^{j+1} \subset \dots \quad (1)$$

The subspace V^{j+1} can define finer details than the subspace V^j . That means, elements in V^j are also elements of V^{j+1} , but V^{j+1} may contain important information not contained in V^j , the finer details. We can use a scaling function and its dilation and translation to construct such a multiresolution analysis

$$\phi_i^j(x) = \sqrt{2^j} \phi(2^j x - i), \quad i = 0, \dots, 2^j - 1 \quad (2)$$

What we use is the simplest Haar wavelet. Its 1-D scaling function is defined as

$$\phi(x) = \begin{cases} 1, & 0 \leq x < 1 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Next, to describe the finer details, we can define subspace W^j which is the orthogonal complement of two consecutive subspaces V^j and V^{j+1} . That is, $V^{j+1} = V^j \oplus W^j$. W^j is what we called wavelet subspace and it's the subspace of "details" in increasing refinements from V^j to V^{j+1} . The wavelet space is spanned by basis functions, as follows:

$$\psi_i^j(x) = \sqrt{2^j} \psi(2^j x - i), \quad i = 0, \dots, 2^j - 1 \quad (4)$$

These functions are the so-called wavelets and they can better characterize important features of a signal or a function than scaling functions do. The corresponding 1-D Haar wavelet is

$$\psi(x) = \begin{cases} 1, & 0 \leq x < \frac{1}{2} \\ -1, & \frac{1}{2} \leq x < 1 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

2-D Haar wavelet is a natural generalization of 1-D Haar wavelet. For image processing, a standard 2-D Haar wavelet transform can be implemented as 1-D Haar

wavelet transform applied on rows of the image followed by another 1-D Haar transform applied on the columns of the transformed image. A 2-D Haar wavelet transform of an image produces a coarse approximation image and three details images. If necessary the approximation image can be further transformed similarly. Starting from the coarsest approximation image, these approximation images define the nested subspaces V^j , ($j=0, 1, 2, \dots$). Similarly those details images define the wavelet spaces W^j ($j=0, 1, 2, \dots$). Figure 3 shows the Haar transform of an image containing a trash can (colour, contrast and brightness have been changed in order make more visible the result of the transform).



Figure 3. A 1-level Haar wavelet transform

In this work, a 3-level Haar wavelet transform is used. This means that the approximation band of the first transform is transformed again, and the same procedure is repeated once more. There are totally 10 subbands produced from one image after transform. The final set of bands is illustrated in Figure 4.

0	1	4	7
2	3		
5	6	8	

Figure 4. Hierarchical subband organization after a 3-level Haar wavelet transform

4 Feature Extraction

In the reported experiments, we define some simple intra-band features, namely *mean value*, *subband energy* and *subband variance*, for each subband. So the feature vector extracted from all subbands contains 30 elements.

Mean Value μ^i ($0 \leq i \leq 9$) is the average amplitude of all coefficients in the i^{th} subband. It can be calculated as

$$\mu^i = \frac{1}{N_i} \sum_{j=1}^{N_i} w_j^i \quad (6)$$

Subband Energy e^i is the average absolute value of all coefficients in the i^{th} subband. It's a rough measure of energy in each subband after wavelet transform. It can be calculated as

$$e^i = \frac{1}{N_i} \sum_{j=1}^{N_i} |w_j^i| \quad (7)$$

Subband Variance v^i is the standard variance of all the coefficients in the i^{th} subband. It can be calculated as

$$v^i = \frac{1}{N_i} \sum_{j=1}^{N_i} (w_j^i - \mu^i)^2 \quad (8)$$

where N_i is the number of coefficients and w_j^i is coefficient in i^{th} subband. As an example, Table I presents these intra-band features of Figure 2.a after the 3 level Haar transform.

The feature vector extracted from the Haar wavelet space has 30 elements from all those 10 subbands. What we can find from table I is that wavelets can provide a sparse representation for images and most energy of the image concentrates in few coefficients having large magnitudes. Furthermore, as we can find, in some subbands the mean values are near zero or very small but their subband energies and variances are relatively large. This implies that pure amplitudes of coefficients may not be right to represent images. Statistical features such as mean values, energies and variances should represent images better than pure coefficients do.

Subband	μ	e	v
0	1249.65	1249.6	85719.3
1	7.89	31.15	3946.16
2	0.50	44.08	7328.83
3	-0.09	13.88	774.03
4	3.34	12.96	883.08
5	3.29	18.22	1616.83
6	-0.09	6.46	116.98
7	1.17	5.99	850.33
8	2.36	9.16	589.12
9	-0.24	4.52	53.76

Table I- the features of figure 2.c after a 3 level Haar transform (including all 10 subbands)

5 Experimental Setup and Results

The experiments reported here are preliminary experiments concerned with recognizing indoor objects such as “person”, “trash can” and “triangle sign”, and therefore enable grounding the corresponding natural

language expressions the robot may encounter in robot-human interaction. These experiments are similar to our previous work [8]. The main difference is the use of features extracted from Haar wavelet space described here. The following is a brief description of the experimental setups and results respectively.

	Yes	No	%Yes
Person	52	104	33
Trash-can	39	117	25
Triangle sign	32	124	21

Table II – Data set composition

A set of 156 images was collected for the experiments in different times but in the same room. Each image can contain none or several instances of those concepts, in arbitrary positions and orientations in a normal office environment. Instances of persons are, actually, feet of persons, because of the hardware limit. Table II shows the composition of the data set we used.

For simplicity and repeatability of the experiment, a client simulator program was used instead of the central manager of Carl. In the learning session, a pre-defined number of images are only used to learn some initial knowledge. For each of the first 40 images, the following is done:

- Apply the Haar wavelet transform to the image;
- Extract feature vector from the Haar wavelet space;
- Ask the teacher if the image contains a person;
- Ask the teacher if the image contains a trash can;
- Ask the teacher if the image contains a triangle sign;
- Send the feature vector with the correspondent yes/no feedback of the teacher as new example to LLL.

The remaining images are used both for training and evaluation. For each image after the 40-th, the currently learned knowledge is firstly applied to it in order to detect persons, trash cans and triangle signs in it. The prediction of LLL is compared to the answers from the teacher and success or failure is recorded. Then, the feedback and the extracted features are sent to LLL to be used as new training example. This training and evaluation procedure was repeated for increasing time intervals between consecutive examples. These intervals are also covered by the delay of human-robot natural language interaction. When classifying an unseen image, both the best MLP obtained so far and the current MLP folds of cross-validation are used. In the second case, the classification is determined by simple voting of the MLP folds. Table III shows average results for the three objects obtained by using the coefficients in the LL subband of the third level Haar transform (band 0 in Figure 4). All 300 coefficients in this subband are used. Results are given for increasing time intervals between consecutive examples sent to LLL. Table IV shows the average results for the three objects

using the feature extraction approach described in section 4, which produces only 30 features.

Average time Intervals (seconds)	Accuracy of folds voting%	Accuracy of best MLP%
5.0	84	82
7.5	84	81
10	86	82
15	84	84
25	84	83
35	86	83
45	85	83

Table III – Average accuracy results obtained by using all 300 coefficients in band 0 of the 3-level Haar transform

As can be seen, in the case of all coefficients in band 0 being used, there is a slight improvement (around 1%) as the average time between examples increases. This is understandable since it results in a longer network training time. However, in the case of intra-band features, the top performance is achieved more or less independently of training time. This is also expected, since the fact that only a small number of features are used allows the learning algorithm to converge faster.

Average time Intervals (seconds)	Accuracy of folds voting%	Accuracy of best MLP%
5.0	88	86
7.5	87	87
10	87	85
15	89	86
25	88	86
35	88	87
45	86	86

Table IV – Average accuracy results in case of intra-band wavelet feature based strategy

Table V compares the global averages of the accuracies for each approach, including the Blocked DCT approach, as reported in a previous publication [8]. In both wavelet-based approaches, we notice that the performances are 2% to 5% better than the Blocked DCT based approach. It can also be seen that the folds voting leads to an accuracy of 2% higher on average than the best MLP. Moreover, the performance of the simple intra-band feature-based approach outperforms in 3% the other wavelet-based approach, in which all coefficients of band 0 are used. This is true both for folds voting and best MLP.

Approach	Accuracy of folds voting%	Accuracy of best MLP%
Blocked DCT (300 coefficients)	83	81
Band 0 (300 coefficients)	85	83
Feature extraction	88	86

Table V – Comparison of the two approaches

Figures 5, 6, and 7 show the evolution of the performance over the three problems as previously unseen images are classified and then added to the server database. The x-axis in these three figures is the number of samples. Here it goes up to 116. These diagrams are taken from the third experiment (third line of Table IV). It can be clearly seen, especially from Figure 5 and Figure 6, that the incremental introduction of new samples can lead to the steadily increase of the accuracy of visual object recognition. These results are quite acceptable given the fact that the collected images contain the objects in a great variety of positions and scales.

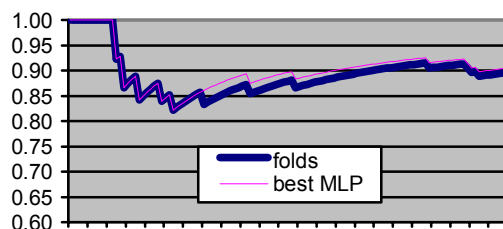


Figure 5. Evolution of recognition accuracy for "person"

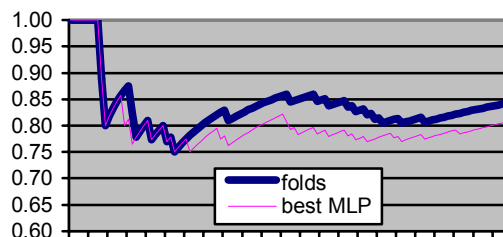


Figure 6. Evolution of accuracy for "trash-can"

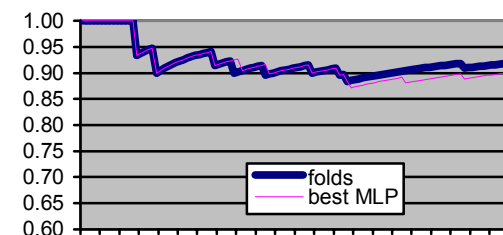


Figure 7. Evolution of accuracy for "triangle sign"

6 Discussion and future work

One issue that should be considered in online learning research is how to evaluate the performance of the learning algorithm. Traditionally, a (one-step) learning algorithm is evaluated using a part of the sample set that

is not used in its training session. But for a lifelong learning algorithm this is very difficult because: i) when the learning session starts, it can't be interrupted; ii) the samples are collected online, so it's not possible to pre-divide the samples for training and evaluating respectively. So we use some predefined number of samples only used for training, and after this, each sample is firstly evaluated by currently learned knowledge and then used as new training sample.

Another issue is about the simple statistical features we used. Although the experimental results are positive, we can't ensure that these subband features represent all the information about the images correctly. They simply reflect the "intra-subband" information in each subband. Thus some harmful information such as illumination variation and background information, also be integrated into or not efficiently eliminated from the learning session. Furthermore, it cuts off the natural relationships of wavelet coefficients among different subbands if we only utilize the intra-band information. These inter-band relationships within a quad-tree formed by the wavelet coefficients are crucial for discriminating the object classes. Without them, we can't achieve good discrimination among the objects of interest only using intra-band information in object recognition and detection. This is among our future work.

7 Conclusion

A Haar wavelet feature based approach for recognizing indoor objects in context of human-robot interaction was presented. In the reported experiments, concerned with concept grounding through visual object recognition, collected images are first pre-processed by 3-level Haar transform, to extract simple intra-band statistical features which are suitable for online learning, then categorized by a teacher and, finally, sent to the learning server in the training session. The results show that this approach is very promising, particularly for our real-time robotics application. Its main advantage with respect to pure coefficients in LL subband is a much smaller feature vector with comparable classification accuracy.

While this work has focused on recognizing indoor objects using simple intra-band statistical features, future work will progressively address the problem of finding the most informative inter-band information along with intra-band information from all subbands after Haar transform for object recognition. Another aspect that increasingly receives our attention is the detection and location of objects of interest in images, possibly in wavelet domain. Besides simple symbols, future work will also address grounding of more complex language expressions, through visual object recognition under the instruction of human teachers.

Acknowledgements

This work is funded by IEETA, Universidade de Aveiro, Portugal, under a PhD grant to Q.H. Wang.

References

- [1] P. Q. Dinh, C. Dorai and S. Venkatesh, "Video Genre Categorization Using Audio Wavelet Coefficients", *Proc. ACCV2002*, Jan 23-25, Melbourne, Australia.
- [2] C. Garcia, G. Zikos and G. Tziritas, "Wavelet Packet Analysis for Face Recognition", *Image and Vision Computing*, Vol.18(4), February 2000, pp.289-297.
- [3] C. Garcia, G. Tziritas, "Face Detection Using Quantized Skin Color Regions Merging and Wavelet Packet Analysis", *IEEE Transactions on Multimedia*, 1(3), September 1999, p.264-277.
- [4] S. Harnad, "The Symbol Grounding Problem", *Physica D*, vol. 42, pp. 335-346, 1990.
- [5] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7): 674-93, 1989.
- [6] C. P. Papageorgiou, M. Oren, and T. Poggio, "A General Framework for Object Detection", *Proc. International Conference on Computer Vision*, Bombay, India, 1998.
- [7] L. Seabra Lopes, "Carl: from Situated Activity to Language-Level Interaction and Learning", *Proc. IROS'02*, pp. 890-896, Lausanne, 2002.
- [8] L. Seabra Lopes and Q. H. Wang, "Towards Grounded Human-Robot Communication", *Proc. IEEE International Workshop on Robot-Human Interaction '02*, pp. 312-318, Berlin, Germany, 2002.
- [9] L. Seabra Lopes and A. J. S. Teixeira, "Human-Robot Interaction through Spoken Language Dialogue", *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 528-534, Japan, 2000.
- [10] L. Seabra Lopes and J. H. Connell, "Semisient Robots: Routes to Integrated Intelligence", in [11], pp. 10-14.
- [11] L. Seabra Lopes and J. H. Connell, eds. (2001) *Semisient Robots* (special issue of *IEEE Intelligent Systems*, vol. 16, n. 5), Computer Society.
- [12] E. J. Stollnitz, T. D. DeRose, D. H. Salesin, "Wavelet for Computer Graphics: A Primer I", *IEEE Computer Graphics and Applications*, 15(3):76-84, 1995.
- [13] E. J. Stollnitz, T. D. DeRose, D. H. Salesin, "Wavelet for Computer Graphics: A Primer II", *IEEE Computer Graphics and Applications*, 15(4): 75-85, 1995.
- [14] Z. H. Sun et al, "A Real-time Precrash Vehicle Detection System", *Workshop on Application of Computer Vision*. Orlando, FL, USA, 2002.
- [15] J. Weng, "A Theory for Mentally Developing Robots", *Proc. Int'l Conf. Development and Learning*, Cambridge, MA, USA, 2002.