Proceedings of ICAR 2003
The 11th International Conference on Advanced Robotics
Coimbra, Portugal, June 30 - July 3, 2003

Wavelet-based Indoor Object Recognition through Human Interaction

QingHua Wang **Luís Seabra Lopes**
Departamento de Electrónica e Telecomunicaçoes/IEETA
Universidade de Aveiro, 3810-193, Aveiro, Portugal
qhwang@ieeta.pt lsl@det.ua.pt

Abstract

In this paper a preliminary work towards grounded robot concept learning through its vision system and human interaction is presented. With a lifelong learning server, introduced in previous work [7], the robot can learn to recognize instances of such concepts as "Person", "Trash-can" and "Triangle sign" from the Haar wavelet transform of its vision space based on the instruction of a human teacher. Some experimental results and comparison with our previous work are also presented.

1 Introduction

Automatic object recognition is an important step in our project, CARL [1], 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 Carl, which can execute such tasks as serving food in a reception or acting as a host in an organization [6, 9].

Although so many efforts have been reported to address the problem of automatic object recognition under unconstrained conditions, it's still open for new trials since it's usually crucial premise for problems such as symbol grounding [1] and image understanding. Generally speaking, the difficulties of object recognition lie in three aspects. Firstly, we need a more appropriate image representation than the pixel based representation because it's difficult to represent objects and their relationships directly in terms of pixels. Secondly, images usually have a dimension of several thousands of pixels or more, so it's not easy to directly process them. Most methods only work well in small dimension, so we need dimension reduction before we can use them. And thirdly, usually we can't collect large numbers of samples for training the classifiers. We may collect enough positive samples, but we often can't collect enough negative samples for training since there are too many objects not belonging to the particular object type.

Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are among the most popular methods addressing the first two problems mentioned above in object recognition, especially for face recognition (To the third problem above, bootstrapping usually used). A good survey of face recognition research may be found in [15], most methods described there can certainly be generalized to objects other than faces. What we notice is PCA or LDA is usually applied to images from constrained scenes, usually constrained illumination or background. Almost all existing databases in use, e.g., face database FERET[2], were collected in this style. For real world problems, in which illumination or background varies frequently, we can't benefit from these methods too much. Another issue we must consider is the (near) real time processing requirement. PCA or LDA doesn't fit this requirement since it's usually computationally intensive.

In this paper we present a wavelet-based approach for indoor object recognition, which is the first step of concept learning, in context of human-robot interaction. Wavelet transform is widely and successfully applied in the domains of signal processing and data compression. In the context of object recognition and detection, wavelet transform is believed to be good at localizing edges and other anomalies. In our application wavelet transform not only serves for image representation, but also for dimension reduction. One of the simplest but fastest wavelet transforms, Haar wavelet transform, is used in the work reported below.

There are already several applications successfully applying Haar transform into object detection and recognition. In [5], a general trainable framework for object detection based on an "overcomplete dictionary" of Haar wavelets is presented. Haar wavelet representation of certain object classes (e.g. pedestrians

[1] Project "Communication, Action, Reasoning and Learning in Robotics", FCT PRAXIS/ 12121/98

[2] http://www.nist.gov/srd/

and cars) is learned directly from class instances in a two-stage process. At first, a small subset of the overcomplete dictionary of wavelet basis functions, capturing the characteristics of the specific object class is learned, based on the statistical analysis of the class instances. Then a class model is learned from these selected basis functions by support vector machines (SVMs).

In [12], a real-time precrash vehicle detection system is described. To detect vehicles on road, a hypothesis generation and verification strategy is used. First of all, the vehicle candidates are predicted from multi scales of the same image using some heuristic rules and constraints. Then a 5 level Haar wavelet is applied on these candidates (scaled to predefined size 32×32) and all coefficients except those in the HH subband of the first level transform, total 768 wavelet coefficients, are used as input to SVMs off-line trained in the same style and thus a decision is given.

Similar to this, in [3] a wavelet-based neural network is proposed for moving vehicle detection. First moving regions are located by frame difference. Then a 2 level Haar transform is applied to the moving object candidate. All 1024 (32×32 regions) wavelet coefficients are used as inputs to a neural network.

The rest of this paper is organized as follows. The wavelet theory is briefly presented in Section 2. The proposed approach is described in Sections 3 and 4. Section 5 presents 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 Haar Wavelet

Although there is specialized literature describing wavelet transformation [4, 11], 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:

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

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

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

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

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

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:

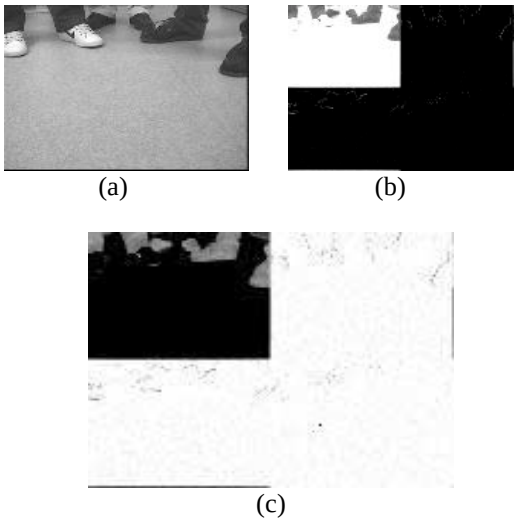
$$\psi_i^j(x) = \sqrt{2^j}\,\psi(2^j x - i), \quad i=0,\ldots,2^j-1 \quad (4)$$

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

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

Since the Haar basis is an orthogonal basis, its transform only provides non-redundant representation of the signal. The 2-D Haar wavelet is an extension of 1-D Haar wavelet. For image processing, standard 2-D Haar wavelet transform can be implemented as 1-D Haar wavelet transform applied on rows of image followed by another 1-D Haar transform applied on the columns of the transformed image. Below is an original image sample used in our work and its 1 level Haar transform respectively (Figure 1.b). As we can find, LL (Low-Low, top-left band) subband keeps most information of the original image. It's an approximation of the original one. The LH (Low-High, bottom-left) subband detects fine details in vertical direction. The HL (High-Low, top-right) detects fine details in horizontal direction. Information in these two subbands is also crucial. Usually information in HH (High-High) subband is noise.

In our real application, a 3 level Haar transform (applied to the original image and then twice to the LL subband recursively) is used. So there are totally 10 subbands after transform. Table I presents some statistical information of Fig 1.a after a 3 level Haar transform.



(a) (b)

(c)

Figure 1: (a) original image; (b) its corresponding 1 level Haar transform; (c) is the invert and magnification of (b) to show the edges in LH, HL and HH subbands.

Table I: the statistical information of Fig 1.a after a 3 level Haar transform (including all 10 subbands)

Band	Max. Coefficient	Min. Coefficient	Mean	σ^2
0	1821.75	419.25	1249.5	85719.3
1	263.13	-336.38	7.89	3946.05
2	375.2	-435.38	0.50	7328.83
3	158.50	-193.13	-0.09	774.04
4	162.25	-305.50	3.34	883.07
5	230.00	-219.75	3.29	1616.83
6	103.25	-76.00	-0.092	116.97
7	102.00	-93.00	1.17	150.00
8	183.00	-147.50	2.36	589.12
9	77.00	-117.00	-0.24	53.76

What we can find from table I is that, after Haar transform, most coefficients are near zero and most energy of the image concentrates in few coefficients having large magnitudes. So wavelets can provide a sparse representation for images and this is why image compression works in the wavelet domain.

3 The Proposed Approach

Generally speaking, robot learning should be seen as a lifelong process [8, 7]. The focus is to assign the robots a kind of capability to self-development with rich sensory-motor skills and minimal initial knowledge, rather than to assign robots rich knowledge in advance. For more details in this direction see [10, 14]. After this construction phase is concluded, a lifelong learning process can start. The robots can learn new skills, explore their environments themselves or under the guidance from human interactants. On-line lifelong learning in robotics has seldom been described. Moreover, the few known systems demonstrating on-line learning capability, still are mostly limited to sub-symbolic 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 communication. Currently, the focus is on recognizing indoor objects, such as "person", "trash can" or "triangle sign". The figure below provides the overview of the approach.

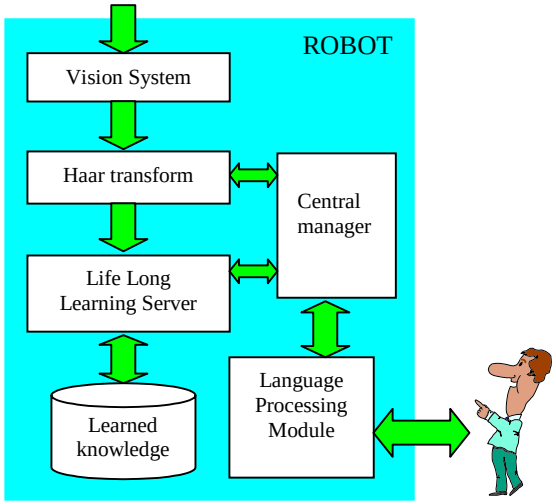


Figure 2. Approach Overview

(a) (b) (c)

Figure 3. Sample image of "Person", "Trash Can" and "Triangle Sign" respectively. These images are re-sized here.

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 [8]. The idea is that, when Carl meets an obstacle, he may ask:

"*Is this a person?*"

Based on the obtained answer ("yes" or "no") and the visual feedback, Carl may store a labeled example. A collection of labeled examples like this will enable a supervised learning algorithm 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, which is briefly described in section 4. A 3 level Haar transform is used to select a most informative subspace for the learning server in order to meet the needs of online learning without sacrificing too much performance. Figure 3 shows some examples used in the training session.

4 Lifelong Learning Server

We have developed a server (named LLL – Lifelong Learning Server [7]) that supports incremental concept learning from examples under the instruction of a teacher. In the real situation, the client manager will be the central manager of Carl. For the experiments reported below, a client program (simulating the central manager) communicates with LLL in a style listed on Figure 4.

Initialize the Client-Server data structures;
Create learning problems according to parameters from client;
In Loop
  Collect sample;
  Collect corresponding instructor feedback;
  If number of examples greater than predefined threshold
    Classify by LLL using current learned knowledge;
    Compare with teacher feedback;
  Send sample and feedback as new training example to LLL;
Until termination;

Figure 4. Client/LLL interaction description

LLL is based on feed forward multi-layer perceptrons. Training is based on the back-propagation algorithm. Weight decay is used to increase the generality of the learned networks (avoid overfiting) in the case of the small sample sets. Bootstrapping is also used to increase the generality. At first, the pool of training examples in LLL is empty and LLL knows nothing about the concepts it is going to learn. During the learning session, LLL maintains a sample pool for further learning. In each learning iteration, LLL randomly selects a sample from the pool for training.

Cross-validation is used for online performance evaluation. When the cross-validation error gets to a new minimum, the main network learned from all examples is saved as the *best network* so far. When evaluating a new case, LLL keeps both the result of applying the best network and the combined results of the folds networks (simple voting is used in classification problems). For more details please refer to our previous work [7, 13].

5 Experiments

The experiments reported here are concerned with recognizing indoor objects such as “person”, “trash can” and “triangle sign”, and therefore enable grounding of corresponding natural language expressions. The following is a description of the experimental setups and obtained results.

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 the camera of Carl is placed at about 25 cm above the floor level. Table II shows the composition of the data set we used.

Table II – Data set composition

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

Each image is transformed using a 3 level Haar transform. The information we use for further learning and evaluation is composed only by the LL subband of the third level Haar transformation. There are totally 300 coefficients in this subband. It is shown in table I that most of the information of the original image space is kept in this subband.

The client simulator program randomly picks images from the pool of 156 previously collected images, until all images have been used. In the learning session, the first randomly selected images are only used to learn some initial knowledge. For each of the first 40 images, the following is done:

-Apply Haar wavelet to the image and selecting the LL subband of the third level;
-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 selected Haar wavelet subspace 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 applied to it in order to detect persons, trash cans and triangle signs. The process is similar to the process above. The only difference is that the results of LLL are compared to the answers from the teacher. Success or failure of LLL's prediction is recorded. Then, also these images are sent LLL to be used as new training examples.

This training and evaluation procedure was repeated for increasing time intervals between consecutive examples. When classifying an unseen image, both the best MLP obtained so far (according to cross-validation) and the current MLPs of the cross-validation folds are used. In the second case, the classification is determined by simple voting of the MLP folds. Table III shows results of our previous work done by using Blocked DCT [7] to extract the most informative data of the sample images where 300 DC coefficients are kept. Table IV shows the average results for the three objects. We see that the folds voting leads to an accuracy of 2% higher on average than the best MLP. We also see a slight improvement as the average time between examples increases. This is understandable since it results in a longer network training time.

Table III – Average accuracy results in case of Blocked DCT based strategy (cited from [7])

Average time interval (seconds)	Accuracy of folds voting%	Accuracy of best net%
7.5	83	77
15	82	78
25	84	82
35	86	84
45	85	82
Global average	84	81

Table IV – Average accuracy results in case of Haar wavelet based strategy

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

As we can find comparing table III with table IV, the performance of the Haar transform based strategy outperforms the Blocked DCT based strategy. For folds voting, Haar based strategy is just 1% higher than DCT based strategy; but for the Best MLP kept, Haar based strategy is 2% higher which means for single classifier Haar based strategy is very promising. This is what we're glad to see. Another advantage of Haar transform is that it's much faster than Blocked DCT. Table V shows average time for extracting information in one image. As we can find, Haar transform is about 30 time faster than blocked DCT with a computer having a 450 MHz CPU and 256 MB RAM. This result is not surprising. Generally speaking, wavelet transform has a time complexity of $O(n)$ [4], and the fastest DCT transform, Fast Cosine Transform, has a time complexity of $O(n \cdot log(n))$ [2], where n is the number of pixels to process. So for near real-time applications, wavelet transform is a very attracting choice.

Table V: Average processing time for the two methods

Method	Average time (ms)
Blocked DCT	500
Haar transform	15

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 (recall that collection contains 156 but a part of it, 40, is used only for training). These diagrams are taken from the fourth experiment (fourth line of Table IV). It can be clearly, seen especially in Figure 5 and Figure 6, that the incremental introduction of false samples can lead to the steadily increase of the accuracy of visual object recognition.

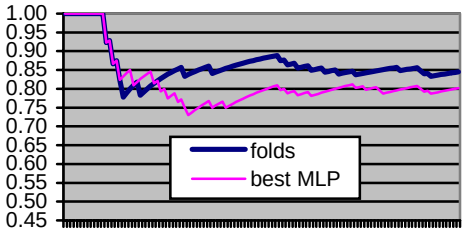


Figure 5. Evolution of recognition accuracy for "person"

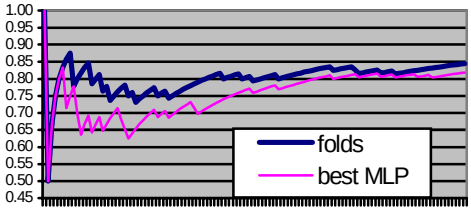


Figure 6. Evolution of accuracy for "trash-can"

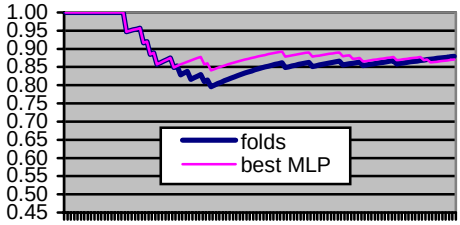


Figure 7. Evolution of accuracy for "triangle sign"

These results are quite acceptable given the fact that the collected images contain the objects in a great variety of positions and scales. At this moment we don't adopt a strategy similar to [3, 12], which detects the candidates of objects of interest first, and then classifies them, since object detection itself is very difficult. Of course this is one direction in our future work.

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 if it is not classified correctly.

Another issue is about the learning algorithm itself. In our approach we use MLPs with weight decay as the basic learning algorithm. One advantage of MLPs is their incremental learning ability with new samples. Another advantage is they can easily find the non-linear decision plane over the complexity data space. The fact is, with a small set of samples in our experiments, the LLL can learn fast and efficiently. But when the number of samples increased to more than 300 the learning session became slow. Future work will, therefore, address the problem of deciding which examples are useful to store.

The third issue is about the LL subband we used. Even if most information concentrates in LL subband after Haar transform, we can't ensure that all the information within this subspace is totally useful for the concept learning. The information within this subspace only keeps the major dimension of the original image space. For example, the background information or illumination information is also kept after Haar transform but this kind of information is harmful for object classification during the lifelong learning session. That's to say, we can't achieve good discrimination among the objects of interest using this kind of information. Some information in other subbands should be considered, just the same as what the image compression does. This is our future work.

7 Conclusion

A Haar transform based approach for recognizing indoor objects from visual information in context of human-robot interaction was presented. It is based on a "lifelong online learning server" and depends on the participation of the human user as a teacher. In the reported experiments, concerned with concept grounding through visual object recognition, collected images are first pre-processed by 3 level Haar transform, to select an informative subspace, suitable for online learning, then classified by a teacher and, finally, sent to the learning server. The results show that the approach based on the Haar transform is very promising, particularly for our real-time robotics application. The main advantage with respect to Blocked DCT is a much higher processing speed with comparable classification accuracy.

While this work has focused on recognizing indoor objects using information of LL subband, future work will progressively address the problem of finding the most informative information from all subbands after Haar transform for object recognition. Another aspect that will increasingly receive 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 instructing of human teachers.

Acknowledgements

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

References

[1] S. Harnad, "The Symbol Grounding Problem", *Physica D*, vol. 42, pp. 335-346, 1990.

[2] F. A Kamangar, and K. R. Rao, "Fast Algorithms for the 2-D Discrete Cosine Transform", *IEEE Transactions on Computers*, Vol. 31(9), pp. 899-906, 1982.

[3] J. B. Kim et al, "A Real-Time Moving Object Detection for Video Monitoring System", *Intl. Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC'01)*, Jul, Japan, Vol. 1, pp. 454-457, 2001.

[4] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation", *IEEE Transactions on PAMI*, 11(7):674-93, 1989.

[5] C. P. Papageorgiou, M. Oren, and T. Poggio, "A General Framework for Object Detection", *Proc. ICCV*, Bombay, India, 1998.

[6] L. Seabra Lopes, "Carl: from Situated Activity to Language-Level Interaction and Learning", *Proc. IROS'02*, pp. 890-896, Lausanne, 2002.

[7] L. Seabra Lopes and Q. H. Wang, "Towards Grounded Human-Robot Communication", *Proc. ROMAN´02*, pp. 312-318, Berlin, Germany, 2002.

[8] L. Seabra Lopes and A. J. S. Teixeira, "Human-Robot Interaction through Spoken Language Dialogue", *Proceedings IEEE/RSJ IROS*, pp. 528-534, Japan, 2000.

[9] L. Seabra Lopes and J. H. Connell, "Semisentient Robots: Routes to Integrated Intelligence", in [10], pp. 10-14.

[10] L. Seabra Lopes and J. H. Connell, edrs. (2001) *Semisentient Robots* (special issue of *IEEE Intelligent Systems*, vol. 16, no. 5).

[11] E. J. Stollnitz, T. D. DeRose, D. H. Salesin, "Wavelet for Computer Graphics: A Primer", *IEEE Trans. CGA*, 15(3):76-84, and 15(4): 75-85, 1995.

[12] Z. H. Sun et al, "A Real-time Precrash Vehicle Detection System", *Workshop on Application of Computer Vision*. Orlando, FL, USA, 2002.

[13] Q. H. Wang and L. Seabra Lopes, "A DCT-based Feature Transformation Strategy for Fast Object Recognition", *Proc. RECPAD'02*, Portugal, 2002.

[14] J. Weng, " A Theory for Mentally Developing Robots", *Proc. Int'l Conf. Development and Learning*, Cambridge, MA, USA, 2002.

[15] M. H. Yang, D. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey", *IEEE Trans. PAMI*, vol. 24, no. 1, pp. 34-58, 2002.