The Tower of Knowledge: a novel architecture for organising knowledge combining logic and probability

Maria Petrou

Communications and Signal Processing Group,
Electrical and Electronic Engineering Department,
Imperial College,
London SW7 2AZ, UK

Abstract. This paper proposes an architecture, called "tower of knowledge", according to which knowledge may be organised in the form of layers of nouns, verbs, adjectives and sensors. A scheme of interpreting scenes using the tower of knowledge and aspects of utility theory is also proposed. The scheme combines concepts of logic approaches with probability theory to propose a method for object recognition and scene labelling.

Keywords: learning, knowledge representation, object recognition, scene labelling

1 Introduction

Classical pattern recognition methods rely on the use of training examples to span the feature space and thus identify the boundaries between the classes they wish to recognise. Thus, they follow the statistical school of thought in relation to learning. However, it is generally acknowledged that learning implies the ability to generalise, and this implies the use of a different methodology for learning. Even neural network based methods, in order to generalise well, rely heavily on the availability of enough training samples to populate adequately the feature space. The training patterns are used by the neural networks to approximate the class boundaries in the feature space with piece-wise linear segments. When an unknown pattern arrives, it can be associated with the class that has been identified to populate the part of the feature space where the pattern appears. Some old [3] and some more recently developed methods [1], that can work with fewer training patterns than straightforward methods, do so by selecting the patterns that matter most in defining the class boundaries, rather than by using some higher level generalisation abilities of the classifier [19]. So, neural networks and pattern classification methods do not really involve mechanisms for generalising beyond the examples that already have been encountered. Statistical learning is slow, as one has to encounter many examples in order to be able to learn the generic characteristics of classes. It is very likely that humans learn

Dagstuhl Seminar Proceedings 08091
Logic and Probability for Scene Interpretation
http://drops.dagstuhl.de/opus/volltexte/2008/1606

this way particularly in early stages of their life. The observed fast learning, even from single examples, often exhibited by humans, must be happening in a different way. It may be attributed to the application of rules of logic which either have already been extracted from the observation of many examples, or have been taught to the learner by some teacher in a ready to use form.

We may conclude, therefore, that we have true generalisation capabilities, only when what is learnt by training examples are rules on how to extract the identity of objects and not the classes of objects directly. If such learning has taken place, totally unknown objects may be interpreted correctly, even in the absence of any previously seen examples.

This conclusion implies that what we have to teach the computer, in order to construct a cognitive system, are relations rather than facts. For example, memorising the dictionary of a language, does not teach a person the language. The person has to learn the relations between words in order to master the language. The relations may be learnt by observing hundreds of examples, i.e. in a statistical way, or they may be inserted to the computer in the form of relational rules. This is in agreement with Winstone's pioneering experiments on teaching the computer to recognise arches. He did not show to the computer all possible types of arch it may encounter. He showed it some examples and counter examples of arches and taught it to recognise relations between components, such as "supports" or "is supported by" [22].

So, both approaches of learning, namely statistical and logic-based, have their role to play in a cognitive system, and they may be emulated when trying to solve problems of recognition in computer vision. An important factor that influences significantly the reasoning process is the way these rules are stored and retrieved. It is not easy to disassociate the reasoning process used, from the way knowledge is stored and information is encoded. This paper presents a scheme of encoding the rules of logic and probabilistic reasoning in a unified framework for recognising objects not only on the basis of how they look like, but also on the basis on what they are used for. Thus, temporal and static appearance information may be incorporated as well as direct communication between sensors and components of the reasoning system.

2 Objects as spatio-temporal entities

Objects exist in space and time. Their existence in time is manifested by the way they are used, i.e. by observed actions involving them. So, when we recognise some object as being a particular type of object, we utilise not only the knowledge we acquired about this object by simply seeing static versions of it, but also by seeing it being used, or even by manipulating it with our hands. We may envisage the following fragment of conversation between a teacher and a learner:
"What is this?"
"This is a window."
"Why?"
"Because it lets the light in and allows the people to look out."

"How?"
"By having an opening at eye level."
"Does it really?"
Such an exchange may be thought of as representing the steps of the reasoning that happens in somebody's head during the process of object identification. We shall use it to identify the components of the system architecture we propose for an artificial cognitive system.

The basic architecture of the proposed system is schematically shown in figure 1. This figure proposes that knowledge may be represented by a series of networks, forming a complex structure that I call the "tower of knowledge". The network of nouns is a network of object names, labels, e.g. "window", "chimney", "door", etc. The network of verbs or actions, is a network of functionalities, e.g. "to look out", "to enter", "to exit", etc. The network of appearances is a network of basic characteristics necessary for a functionality to be fulfilled, e.g. "it is an opening of human size at floor level". So, the flow of knowledge goes like the fragment of conversation given above. The loop closes when we confirm that the object we are looking at has the right characteristics for its functional purpose to be fulfilled. Most classical pattern recognition approaches operate only at the level of nouns of the tower of knowledge. They use relations only in the form of relative geometric arrangements of objects, or object co-occurrences in scenes, in order to capture context, rather than relations that are invoked through joint purpose, joint use, or joint involvement in actions, as it is advocated here.

Note that the generic logic model of an object is encoded in the inter-layer connections of the proposed scheme. The generic knowledge of which sensors are useful for which object or functionality is again implicitly encoded in the connections that start from the various different descriptors and go back to the sensors. In such a scheme there is room for involving sensors other than cameras, for example chemical or pressure sensors, if the characteristic necessary for an object to fulfil a certain functionality involves input other than optical. For example, if the label "food" is invoked at the nouns level, and functionality "to be eaten" is invoked in the verbs level, which in turn invokes the descriptor "smells good", an olfactory sensor may be invoked to check whether the object actually has an acceptable smell for something edible. This example shows that sensors and measurements invoked by the descriptors level may be different from those originally observed and fed bottom up to the nouns level to start the process of labelling. An example which makes such a scheme very clear is the case when one tries to discriminate between a realistic looking flower and a real one.

For the tower of knowledge scheme to be implemented in practice, one has to be able to model its various layers of networks and their inter-connections. For this purpose, one has various tools at one's disposal: Markov Random Fields [8], grammars [17], inference rules [20], Bayesian networks [15], Fuzzy inference [23], etc. I would exclude from the beginning any deterministic crisp approaches, either because things are genuinely random in nature (or at least have a significant random component), or because our models and our knowledge is far

too gross and imperfect for creating crisp rules and dogmatic decisions. The most commonly used classical pattern recognition approaches which incorporate contextual information (at the level of nouns only) are Markov Random Fields (MRF) and Bayesian Approaches, in particular Probabilistic Relaxation (PR). I will examine next these two approaches.

3 Markov Random Fields

Contextual influence within the same level of the tower of knowledge often is modelled by a classical MRF, represented by an undirected graph. I argue here that directional MRFs should be used instead, often called belief networks. A directional (or asymmetric) MRF captures better the mutual influence between labels. For example, the label "staircase" may trigger the label "door" more frequently than the label "door" triggers the label "staircase". This asymmetry in the interactions is a manifestation that Markov Random Fields (MRFs) are not applicable here in their usual form in which they are applied in image processing. An example of the interactions in a neighbourhood of an MRF, defined on a grid, is shown in Fig. 2b. This MRF, and the weights it gives for neighbouring interactions, cannot be expressed by a Gibbs joint probability density function [5]. For example, the cell at the centre is influenced by its top left neighbour with weight -1, but itself, being the bottom right neighbour of the cell at the top left, influences it with weight $+1$. This asymmetry leads to instability when one tries to relax such a random field, because local patterns created are not globally consistent (and therefore not expressible by global Gibbs distributions) [16]. According to Li [9, 10, 11], relaxations of such MRFs do not converge, but *oscillate* between several possible states. (Optimisations of Gibbs distributions either converge to the right interpretation, but more often than not, they *hallucinate*, i.e. they settle on wrong interpretations.)

So, one could model the network at each level of the tower of knowledge shown in Fig. 1, using a non-Gibbsian MRF, e.g. a belief network [5]. The interdependencies between layers might also be modelled by such networks, but perhaps it is more appropriate to use Bayesian models, as the inter-layer dependencies are causal or diagnostic, rather than peer-to-peer.

4 Bayesian Inference

Bayesian approaches have been used so far in two ways: either in the form of probabilistic relaxation (PR) [7, 2] or in the form of Pearl-Bayes networks of inference [15]. Probabilistic relaxation has its origins in the seminal work on constraint propagation by Waltz [21], who used crisp constraints and solved once and for all the problem of globally inconsistent labellings that used to lead to impossible objects [6]. Probabilistic relaxation updates the probabilities of various labels of individual objects by taking into consideration contextual information [7]. As this contextual information is in effect peer-to-peer, probabilistic relaxation is **not** an appropriate tool for modelling causal relationships.

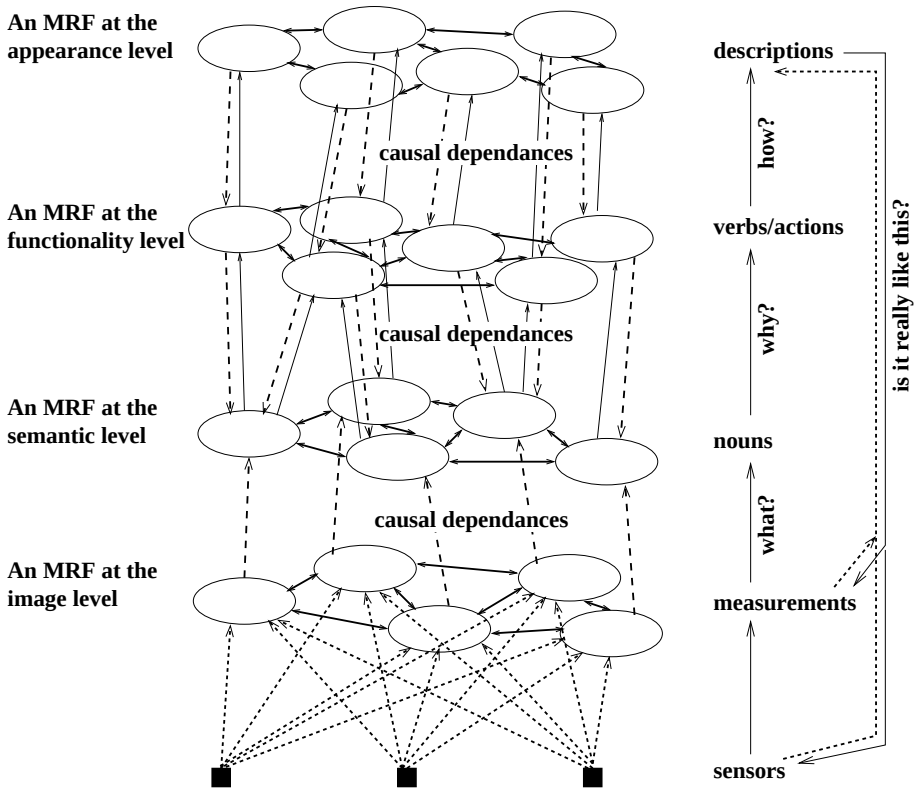

Fig. 1. The tower of knowledge: how knowledge may be organised. The double-headed arrows represent contextual interactions. The thin continuous arrows represent queries. The dashed arrows represent answers, i.e. transfer of information. The level of interest in a cognitive vision task is the level of nouns, where we wish to assign labels to objects. Examples of nodes with contextual connotations in the network of nouns are "door", "window", "balcony". Examples of nodes with contextual connotations in the network of functionality are "lets air in", "lets light in", "allows a person to enter". Examples of nodes with contextual connotations in the network of descriptions are "has a glass pane", "is at eye-level", "has a handle to open it".

It is rather an alternative tool to MRFs discussed in the previous section for modelling influences at the same layer. Probabilistic relaxation, just like MRF relaxation, is not guaranteed to converge to a unique global solution, unless special conditions are obeyed [18]. This, however, is not an issue in reality: labellings of scenes do not have to be globally consistent, but only locally consistent. This statement seems to be in contradiction with the previous statement, saying that probabilistic relaxation is the generalisation of Waltz's algorithm which solved the problem of inconsistent labellings in the 60s. This contradiction, however, is only superficial. The problem of inconsistent labellings of the 60s was referring to the labellings of single solid objects, by labelling their sub-parts [4] and not

the labellings of scenes that contain many different objects, where constraints between objects are far weaker than constraints within the subparts of the same solid object.

The second form of Bayesian approach is that of Pearl-Bayes networks of inference. Here the relations may be causal, and so these networks are appropriate for inter-layer inference. Bayesian approaches depend on conditional probabilities. How to choose these conditional probabilities has always been a problem for such methods. Conditional probabilities may have to be learnt painfully slowly from hundreds of examples. Alternatively, conditional probabilities may be transferred ready from another already trained network: the network of the teacher. This transference is equivalent to choosing them to have some parametric form (e.g. Gaussian) with parameters chosen "arbitrarily". The arbitrary choice of form and parameters usually leads to the criticism of the approach being ad-hoc or unjustified. It is not, if the teacher simply transfers their own hard gained knowledge to the pupil (the computer). Such an approach leads us to new theories, like for example the so called "utility theory" [12].

Utility theory is a decision theory. Assigning labels to objects depicted in an image is a decision. In the Bayesian framework we make this decision by maximising the likelihood of a label given all the information we have. In utility theory, this likelihood has to be ameliorated with a function called "utility function", that expresses subjective preferences or possible consequences of each label we may assign. The utility function multiplied with the Bayesian probability of each label and summed over all possibilities leads in one pass only to the final label. So, this approach avoids the iterations used by MRFs and PR. The utility function may be identified with the innate meta-knowledge somebody has acquired about the world. It is that knowledge, that might have been learnt from many examples, but which now is expressed in the form of conditions and prejudices that cannot be fully justified by the measurements we make. It is the knowledge that tells us to be cautious when we want to buy a car from a man that postponed the appointment we made several times, that did not produce immediately the maintenance record of the car we requested, and so on. Such ideas have been around for some time, without people using the term "utility function". For example, psychologists in the mid-nineties were talking about the so called p-maps and m-maps. The p-maps were meant to be the prior knowledge we have about various possible patterns that we may encounter in life. A p-map guides us to sample a scene more or less carefully at places where it matters or it does not matter, respectively, producing the m-map that is specific to the present situation. One may identify here the p-maps as being the utility functions of today and the m-maps the Bayesian part of labels conditioned on the measurements we have made[1].

In the computer vision context, utility theory has been used by Marengoni [13] to select the features and operators that should be utilised to label aerial

[1] The ideas of p-maps and m-maps first came to my knowledge by Robin Shirley of the Psychology Department of Surrey University, who passed away before he had the chance to make them more concrete and publish them.

images. Further, one may interpret the work of Miller et al. [14] as using a utility function that penalises the unusual transformations that will have to be adopted to transform what is observed to what the computer thinks it is. The authors effectively choose labels by maximising the joint likelihood of the probability density function of the observed transforms and the probability density function of the labels and observations, assuming that transforms and labels/measurements are independent.

5 Modelling the "why" and the "how" in order to answer the "what"

Let us consider the tower of knowledge presented in Fig. 1. We shall formulate here the problem of learning to recognise objects in a scene, using this hierarchical representation of knowledge and utility theory.

Let us assume that we use maximum likelihood to assign labels to a scene. In the conventional way of doing so, object o_i will be assigned label l_j with probability p_{ij}, given by:

$$p_{ij} = p(l_j|m_i)p(m_i) = p(m_i|l_j)p(l_j) \tag{1}$$

where m_i represents all the measurements we have made on object o_i, and $p(m_i)$ and $p(l_j)$ are the prior probabilities of measurements and labels, respectively. Probabilistic relaxation will update these probabilities according to the contextual information received from neighbouring regions. We do not follow that route here. Instead, we shall use the information coming from the other layers of knowledge to moderate this formula. Let us identify the units in the verbs level of Fig. 1 by f_k, and the units at the descriptor level of Fig. 1 by d_l. Then we may choose label l_{j_i} for object o_i as follows:

$$j_i = \arg \max_j \underbrace{\sum_k u_{jk} \sum_l v_{kl} c_{il}}_{\text{utility\_function}(i,j)} p_{ij} \tag{2}$$

where u_{jk} indicates how important is for an object with label l_j to fulfil functionality f_k; v_{kl} indicates how important characteristic d_l is for an object to have the possibility to fulfil functionality f_k, and c_{ik} is the confidence we have that descriptor d_l applies to object o_i.

Note that the value of the utility function expresses the evidence we have that region o_i has the necessary characteristics to fulfil its role as object l_j. For example, if the label we consider of assigning to object o_i is "balcony", the utility function must express whether this object has dimensions big enough to allow a human to stand on it, whether it is attached on a wall, and whether there is a door leading to it. All these are conditions that will allow an object to play the role of a balcony. A learning scheme must be able to learn the values of u_{jk} and v_{kl} either directly from examples (slowly and painfully), or by trusting its teacher, who having learnt those values himself, slowly and painfully over many

years of human life experiences, directly inserts them to the computer learner. The computer learner then must have a tool box of processors of sensory inputs that will allow it to work out the values of c_{il}.

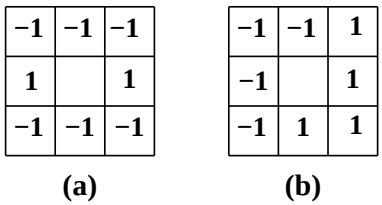

Fig. 2. (a) A local neighbourhood at the pixel level with globally consistent Markov parameters: if this field is relaxed it will lead to horizontal strips of similar labels which will be distinct from the labels above and below. In image processing it will lead to a texture pattern with strong horizontal directionality. (b) A local neighbourhood at the pixel level with globally inconsistent Markov parameters: the top left pixel tells the central pixel to be different from it; the central pixel, seen as the bottom right neighbour of the top left pixel, tells it to be similar to it.

6 Conclusions

I presented here a scheme that may be used to recognise objects. It implicitly encodes the rules of logic in the form of connections made between objects (nouns), functionalities (verbs), descriptors (adjectives) and sensors, each represented by a separate layer of interconnected nodes. Modelling the relations between the nodes of the same layer may be achieved by using non-Gibbsian Markov random fields (e.g. directed graph models or belief networks) or probabilistic relaxation, while transferring information between the different layers may be performed using a probabilistic approach like Maximum likelihood estimation and utility theory.

I argued that learning is characterised by the ability to generalise, and that this can only be achieved if what is learnt is not the labels of the objects viewed, but the rules according to which these labels are assigned. I have also argued that this meta-knowledge may be transferred to the learner (the computer) directly by the teacher (the human developer), in the form of rules, or in the simplest way, by the human using the parameters of the algorithms according to their personal experience and intuition. This puts me at odds with the majority of the community of reviewers who tend to reject papers on the grounds that the parameters have been chosen ad hoc with no proper explanation: these are the cases of the teacher transplanting to the learner their painstakingly acquired knowledge. The alternative is for the learner each time to acquire this knowledge painfully slowly from thousands of examples.

I also argued that we do not need globally consistent labellings of scenes. Global consistency will never allow us to label correctly the scene painted by Magritte of a train storming out of a fire place, because trains do not come out from fire places! It will never allow the computer to recognise green horses with 5 legs, but we, humans, do. So, what we need is fragments of reality and knowledge. The framework I presented views objects as spatio-temporal entities and allows the incorporation of information coming from temporal observations involving them, static images, user prior knowledge and extra sensors to decide their identity. Being an object-centric scheme, it does not try to find a globally optimal labelling, and as such it is expected to be able to cope with the interpretation of bizarre scenes like those mentioned above.

Acknowledgements: This work was supported by EU grant 027113.

References

1. Cortes, C. and Vapnik, V. N.: Support-Vector Networks. Machine Learning Journal **20** (1995) 273–297.
2. Christmas, W. J., Kittler J. and Petrou, M.: Structural matching in Computer Vision using Probabilistic Relaxation. IEEE Transactions on Pattern Analysis and Machine Intelligence **17** (1995) 749–764.
3. Devijver, P. A. and Kittler, J.: On the edited nearest neighbour rule. Proc. 5th Int. Conf. on Pattern Recognition, (1980) 72–80.
4. Guzman, A.: Computer Recognition of three-dimensional objects in a visual scene. Tech. Rep. MAC-TR-59, AI Laboratory, MIT (1968).
5. Heesch, D. and Petrou, M.: Non-Gibbsian Markov Random Fields for object recognition. The British Machine Vision Conference (submitted, (2007).
6. Huffman, D.A.: Impossible Objects as Nonsense Sentences, Machine Intelligence **6** (1971) 295–323.
7. Hummel, R. A. and Zucker, S. W.: One the foundations of relaxation labelling process. IEEE Transactions PAMI **5** (1983) 267–287.
8. Kindermann, R. and Snell, J. L.: Markov Random Fields and their Applications. First book of the AMS soft-cover series in Contemporary Mathematics, American Mathematical Society (1980).
9. Li, Z.: A neural model of contour integration in the primary visual cortex. Neural Computation **10** (1998) 903–940.
10. Li, Z.: Visual segmentation by contextual influences via intra-cortical interactions in the primary visual cortex. Networks:Computation in Neural Systems **10** 187–212.
11. Li, Z.: Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex. Neural Computation **13** (2001) 1749–1780.
12. Lindley, D. V.: Making Decisions. John Wiley (1985).
13. Marengoni, M.: Bayesian Networks and Utility Theory for the management of uncertainty and control of algorithms in vision systems. PhD thesis, University of Massachusetts (2002).
14. Miller, E. G., Matsakis, N. E. and Viola, P. A.: Learning from one example through shared densities on transforms. CVPR (2000).
15. Pearl, J.: Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufmann Publishers Inc. (1988).

16. Petrou, M. and Garcia Sevilla, P.: Image Processing, Dealing with Texture. Wiley. ISBN-13 978-0-470-02628-1 (2006).
17. Schlesinger, B. D. and Hlavac, V.: Ten lectures on Statistical and Structural Pattern Recognition. Kluwer Academic Publishers, Dordrecht, The Netherlands (2002) chapter 10.
18. Stoddart, A. J., Petrou, M. and Kittler, J.: On the foundations of Probabilistic Relaxation with product support. Journal of Mathematical Imaging and Vision **9** (1998) 29–48.
19. Tong, S. and Koller, D.: Support Vector Machine active learning with applications to text classification. Journal of Machine Learning Research **2** (2001) 45–66.
20. Walker, T. C. and Miller R. K.: Expert Systems Handbook: An Assessment of Technology and Applications. Englewood Cliffs, NJ, Prentice-Hall (1990).
21. Waltz, D.: Understanding line drawings of scenes with shadows. The Psychology of Computer Vision, P. Winstone (Ed.) McGraw-Hill (1975) 19–91. (http://www.rci.rutgers.edu/c̃fs/305_html/Gestalt/Waltz2.html)
22. Winston, P. H.: Learning structural descriptions from examples. The psychology of computer vision (1975) 157–209.
23. Zadeh, L. H.: A fuzzy algorithmic approach to the definition of complex or imprecise concepts. Int. J. Man-Machine Studies **8** (1976) 249–291.

This article was processed using the LaTeX macro package with LLNCS style