HTM networks – the future of computer vision systems?

Michal Podpora

Opole University of Technology,
Faculty of Electrical Engineering, Automatic Control and Informatics,
45-272 Opole, Poland
michal.podpora@gmail.com
http://podpora.opole.pl/

Abstract. In this paper author describes a practical implementation of interconnection of an HTM network and a mobile robot's computer vision system. A brief explanation of HTM is followed by detailed description of image processing and compression algorithms. Assuming that all image processing and understanding procedures imitate human eyesight, the proposed image representation can be compressed using the proposed lossy compression methods. This in turn enables the use of computer clusters.

1 Introduction

Today's computer vision systems use sophisticated image processing algorithms and require powerful hardware. Image recognition algorithms give great results but only if the background is fairy clear or recognized objects are well known to the system. Despite the entire technological advance, teaching a computer vision system still differs from "teaching" a human a new object to recognize. A picture is worth a thousand words – and every word describes a variable or property or feature of an object. Extracting these features from an image, interpreting and assigning a proper object are very complex. Image understanding is, indeed, a great challenge for machines.

Unfortunately, not all features of an object are always visible in an image as well as not all of them are always properly determined. Computer system should also be able to discover more than one object in a scene. As John McCarthy said, "1001 words is worth more than a picture" – the more features and objects a system "knows", the more complex processing and uncertain outcomes become.

Computer vision systems are still too far from our biological vision. It might be the right time for a change.

2 HTM networks

There are some features that distinguish HTM from other approaches: an HTM network processes both spatial and temporal information, a trained HTM can reconstruct (generate) missing (spatial or temporal) information.

HTM networks are hierarchical, but their nodes are not connected in a predefined way. Developer can pick predefined nodes for every layer and link them to form any hierarchical structure. It is also possible to design a new node from scratch. The most important node types are sensor nodes, which can be considered as inputs of HTM network. HTM network can process spatial data patterns as well as temporal ones.

2.1 Hierarchical

HTM networks are meant to be the most possible precise mapping of the human brain “algorithm”. The neocortex consists of six layers – their task is to generate new levels of abstraction (Fig. 1). If an object has a tail, whiskers, soft fur and a mouse is lying nearby, the object may be a cat. [1]

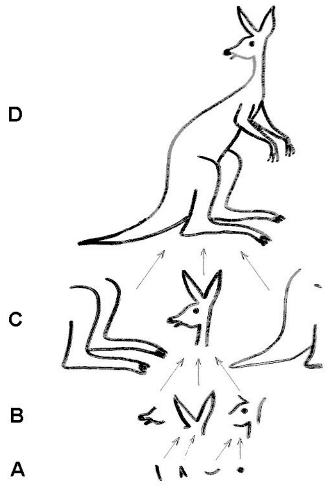

Fig. 1. Levels of abstraction – lower layers generate abstract objects for higher layers of an HTM network

The lowest layers get the information from sensors. This information is processed in higher layer to form another level of abstraction. The sixth layer (the association layer) has less cells, but far more connections – it enables the possibility to link information between various sensors (senses). Every layer has regions – groups of nodes (neurons) identifying a particular feature. Neurons in visual cortex V4 of human brain distinguish certain shapes like blue or red stars [2]; MT cortex detects movement of an object. For actuators, the same algorithm is being used, only the direction of information flow differs (forward connections are switched with feedback connections).

2.2 Temporal

Time is one of the most important aspects of the input data. It can be interpreted simply as another dimension of the input data or as a parameter. A good example of a temporal pattern is a melody [2] – a sequence of sound frequencies. In a melody, the order of sounds should not be changed. However, in most cases the temporal order is not obligatory. In the next chapter saccades will be discussed – rapid eye movements for observing parts of an object. The order of the observed parts is irrelevant.

2.3 Memory

There is an enormous amount of neural feedback connections in the neocortex. Some scientists claim, that more than a half of all of the connections. The main reason for this would be the nature of the learning and remembering algorithms. But there is another reason for this: the brain's "capability of thinking".

The memory of a HTM network is also based on the phenomenon of brain's memory. It is similar to what we know as the autoassociative memory. One of its most important features is the ability to predicate. If not all of the object's features are available it can still classify the object.

2.4 HTM network

Neurons in the lowest layers, regardless of the kind of processed information (sound, vision, touch or other), get the information in the most possible primitive form – as the intensity level of a single parameter (brightness, color, sound frequency, pressure, etc.) Higher layer neurons collect outputs from the sensory layer neurons and become active whenever a specified more complex abstract object is detected (e.g. in sound analysis – from frequencies to phonemes, in vision system – from matrix of pixels to lines and curves).

HTM networks may become valuable in many various applications. For instance, some text recognition (OCR) applications use neural networks and/or dictionaries to avoid recognition errors. Nevertheless, sometimes a dictionary is not enough – if a word contains an erroneous letter and forms other existing word: "flying high as a bind". For a human being this sentence makes no sense. If a machine (or software) is supposed to "understand" our environment or to interact with us, why not design its cognition algorithms basing on our biological algorithms.

3 Human vision

Sight is considered to be the most complex biological sensory input. Analysis and inference is also complicated – four-dimensional, object-oriented, context-dependent data is unquestionably difficult to process. Computer vision systems use video cameras or other acquisition hardware in substitute of human eyes. Seemingly, this is the easiest interconnection between vision and a computer. Unfortunately, the largeness of the input data makes the processing even more difficult. However, in an eye, rods and cones are not equally spaced and they do not form a rectangular matrix. The "resolution" of an eye is higher near so-called yellow spot. Outside the yellow spot the quality of image is relatively poor and, therefore, insufficient for processing or object detection algorithms. However, this great spatial variability of quality is transparent for brain's recognition and inferring algorithms. The main reason for this, are saccades (rapid eye movements that change the spatial input data). The brain composes the final image and infers the meaning of an object in the scene.

4 Computer vision

The majority of all computer vision projects ignore the nature of human vision algorithm and the existence of yellow spot. Usually the whole image is being scanned and processed, except for the Active Vision approach, where only a part of an image is analyzed. Lossy compression methods (and any algorithms negatively affecting the image quality) are not considered during the design and implementation of a computer vision system.

The implementation of a computer vision system that is based on a human vision is possible with lossless compression and maximum quality, but such algorithms seem to be not scalable. Image processing becomes more and more complex, and there is no possibility to use a computer cluster. The distribution of video frames within a computer cluster is very difficult, while the data transmission is the bottleneck of today's supercomputers. However, there is no image distribution of any issue within the brain.

4.1 Point-of-interest (virtual yellow spot)

Designing an HTM-ready vision system is currently possible. A computer vision system made of a webcam, trained HTM network and some sensor nodes is used as one of the basic HTM tutorials [3]. The Vision4 demo application is capable of [4] recognizing four predefined objects (sailboat, rubber duck, cow, cell phone) regardless of color, exact shape or orientation with the use of HTM networks. The Vitamin D Video application [5] in real time detects movement and classifies the moving object as a human being or not. These applications may be great examples of HTM philosophy, but they are still based on "the best-possible-video-quality" approach and, therefore, their vision systems are not scalable.

The practical implementation of virtual yellow spot [6] and saccades [7] remarkably reduces the image (or video stream) data size [8].

The virtual yellow spot defines an image with maximum quality in the neighborhood of specified coordinates. The saccades procedure causes the coordinates of virtual yellow spot to change.

4.2 Image representation

The Point-of-interest conception [6] is based on the JPEG2000's Region-of-interest. While in JPEG2000 the Region-of-interest has a separate threshold value (which causes different quality of the region and the rest of the image), in virtual yellow spot approach the threshold value is variant. It equals zero at the yellow spot's coordinates and rises with the distance from that point [6]. A variable threshold parameter of the JPEG2000's discrete wavelet transform also enables the possibility to use lossy compression methods in conjunction with DWT transformation procedures. [9]

4.3 Saccades (feedback)

In its most primitive form saccades should be implemented just as a feedback method of passing new coordinates to image acquisition algorithms. In a more complicated form, saccades can be implemented as a complex feedback involving physical movement of image acquisition hardware and/or the robot's body.

5 Benefits

Implementing virtual yellow spot combined with saccades significantly reduces the data size of image frames. It turns out, that (to some extent) the visual data can be transferred and processed within a computer cluster. This implies the use of Task Parallelism (a limited set of different cluster nodes processing various aspects of the acquired visual data).

HTM network algorithms are similar to our brain's algorithms – creating machines that use HTM for AI purposes may produce a new, much more "intelligent" class of robots. Transferring the vision processing and inferring algorithms to a computer cluster also gives us new possibilities of the robot's AI development.

References

1. Hawkins, J., Dileep, G.: Hierarchical Temporal Memory – Concepts, Theory and Terminology, available on-line (accessed 2010-07-31): http://www.numenta.com/ /Numenta_HTM_Concepts.pdf
2. Hawkins, J., Blakeslee, S.: On Intelligence, Times Books, 2004

3. Numenta: Sample Vision Networks, available on-line (accessed 2010-07-31): http:// /numenta.com/vision/webservices-embed.php
4. Numenta: Vision4 Demo, available on-line (accessed 2010-07-31): http://numenta.com/ /about-numenta/demoapps.php
5. VitaminD: Vitamin D Video, available on-line (accessed 2010-07-31): http:// /www.vitamindinc.com
6. Podpora, M., Sadecki, J.: Biologically reasoned point-of-interest image compression for mobile robots. In: Hippe, Z., Kulikowski, J.: Human-Computer Systems Interaction – Backgrounds and Applications. Springer-Verlag, Berlin Heidelberg New York (2009) 389–401
7. Podpora, M.: Dynamic re-definition of Region-of-Interest in vision system's feedback. IEEE eXplore (2009)
8. Podpora, M.: Computer vision in parallel computing. ISTET'07, Przeglad Elektrotechniczny, Vol. 11/2007, Szczecin Warszawa (2007)
9. Podpora, M.: Biologically reasoned machine vision: RLE vs. entropy-coding compression of DWT-transformed images, Proceedings of the 14th Conference Student EEICT 2008, Vol.4, Brno (2008) 457–460

Acknowledgement: Work co-financed by European Social Fund