# Invariant Object Recognition using Spiking Neurons

Qian Liu

*Abstract*—The major stumbling problem of the computer object recognition lies in the poor robustness to the transformations. Exploring and mimicking invariant object recognition within the brain is a promising approach to tackling the computational difficulty; in turn it also contributes to the understanding biological visual processing.

Thanks to its high-performance massively parallel processing, SpiNNaker makes it possible to simulate large-scale neural networks in real-time. As a first milestone of this study, a recognition system for dynamic hand postures is developed on a neuromorphic hardware platform. Future work is proposed to build an object recognition system with position, scale and view invariance by modelling the hierarchical visual pathway up to the inferotemporal (IT) cortex. The proposed plan of the system will be able to recognise 200 objects in real time exploiting Integrate-and-Fire(LIF) neurons.

## I. Introduction

### A. What Is Object Recognition?

Object recognition is the process of assigning labels to particular objects, ranging from precise labels ('identification') to coarse labels ('categorisation') [1]. This includes the ability to accomplish these tasks under various identity preserving transformations such as object position, scale, viewing angle, background clutter and etc. (known as transformation invariance). The brain can accurately recognise and categorise objects remarkably quickly, for example object recognition time in monkeys is under 200 ms [2] and the images are presented sequentially in spikes less than 100 ms [3]. This research focuses on this rapid and highly accurate object recognition, 'core recognition', which is defined in [4].

### B. Why Is It Important?

The human brain recognises huge amount of objects rapidly with ease even in cluttered and natural scenes. However, artificial object recognition systems are poorly robust to these various transformations. Each encounter of an object on the retina is unique because of differing illumination (lighting conditions), position (projection locations on the retina), scale (distances and sizes), pose (viewing angles), and clutter (visual contexts). In addition, a difficult specificity-invariance trade-off occurs in the categorisation tasks, since the recognition should be able to discriminate different object classes (intra-class variability) while at the same time remaining tolerant to image transformations.

To solve the computational difficulties of invariant object recognition, we explore and model the visual processing within the brain; in turn, it also unveils the mechanism of biological visual processing by mimicking the neural activity in the visual system. Moreover, energy-efficiency improvements following from the great energy efficiency of biological systems will help in building object recognition systems, e.g. posture recognition for human-machine interfaces in mobile devices.

### C. How to Mimic The Brain?

To explore how brain may recognise objects, we have employed a biologically-inspired Dynamic Video Sensor (DVS) silicon retina [5]. and a SpiNNaker system [6], which is a massive parallel computing platform aimed at real-time simulation of Spiking Neural Networks (SNNs). Thanks to its high-performance processing of large-scale neural networks, we explore biological approaches of visual processing by mimicking the functions of different layers along the ventral visual pathway. The goal of this proposed research is to build an object recognition system with position, scale and view invariance by modelling the hierarchical visual pathway within the brain. The proposed plan of the system will be able to recognise 200 objects in real time exploiting LIF neurons.

Building a real-time recognition system for dynamic hand postures is a first step of exploring visual processing in a biological fashion and is also a validation of the performance of the neuromorphic platform. This preliminary work achieved the first milestone of the research which aims at building a position-invariant object recognition system exploiting V1-like neurons (primary visual cortex) to classify five hand postures.

## II. Biological Aspects

### A. The ventral visual pathway

The ventral visual pathway (Figure 1A) starts from the primary visual cortex V1 in the occipital cortex through areas such as V2 and V4 to the Inferotemporal (IT) cortex.

**Primary Visual Cortex: V1.** As the simplest and earliest cortical area in the ventral stream, the primary visual cortex V1 is the best-studied since the well-known discovery of the orientation selectivity by Hubel and Wiesel [7] in 1958. In the spatial domain, V1 neurons are tuned to Gabor-like transforms applied to their small local receptive field. In theory, these Gabor-like filters together can carry out neuronal processing of spatial frequency, orientation, motion, direction, speed, and many other spatio-temporal features.

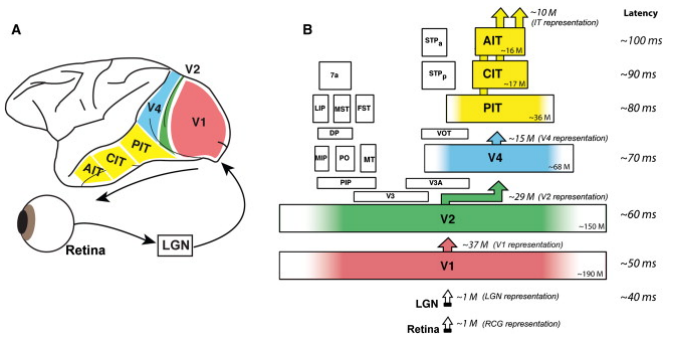



Fig. 1: The ventral visual pathway and abstraction layers [1].

**Visual Areas V2/V4.** The responses of many V2 neurons are also modulated for complex properties: orientation of illusory contours [8], binocular disparity [9], and whether the stimulus is part of the figure or the ground [10]. Although V4 is mainly modulated for colour recognition, it is also tuned for orientation and spatial frequency similar to V1. Comparing to V1, V4 responds to more complex object features with intermediate complexity.

**Inferotemporal Cortex: IT.** The complexity increases along the ventral stream towards anterior IT (AIT) where objects are represented and recognised [11]. The high-order complex features includes the combinations of colour or texture with complicated shapes [12], and body parts such as faces and hands [13]. The distinguishing features of the IT cortex is that the neuronal responses are position and size invariant [14], and also invariant to changes in luminance, texture, and relative motion [15].

**Hierarchical Feed-forward Organisation.** The corresponding hierarchical organisation is showed in Figure 1B. Each area is plotted with the size proportional to its cortical surface size. Approximate total number of neurons of both hemispheres is shown in the corner of the cortical areas. The approximate number of projections is written above each block. In addition, the colour dedicates to the processing of central 10° of the visual field. At last, approximate median response latency is listed on the right.

### B. Object Representation in IT

The neuronal representation in the cortical area of IT is considered to be the spatio-temporal pattern of spikes. The spiking activities of single neurons and populations are thought to hold the key to encode visual information.

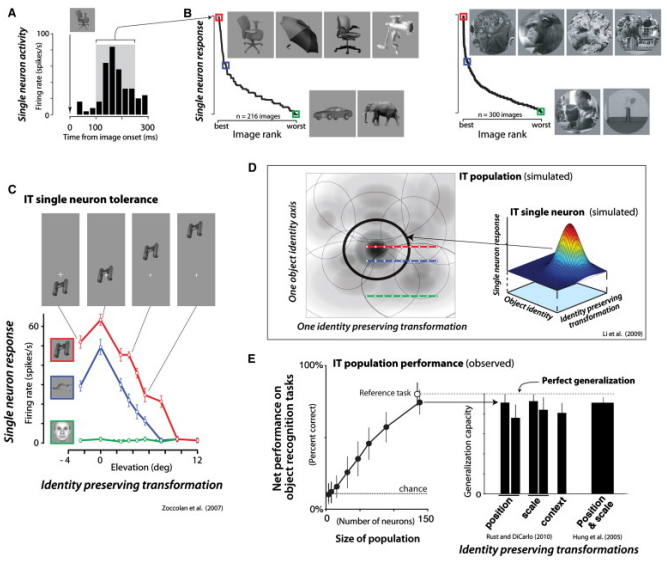



Fig. 2: IT single-neuron properties and their relationship to population performance [1].

Figure 2 illustrated typical IT neuronal activity. Figure 2A shows the spike count of a single neuron in time bins of 25 ms for a duration of 300 ms after the presentation of a visual image, and the highlighted 'decoding' window is adjusted to the latency of the conductance along the ventral stream. The spike count of the 'decoding' window is well modulated for object identity (Figure 2B), position (Figure 2C) and etc.

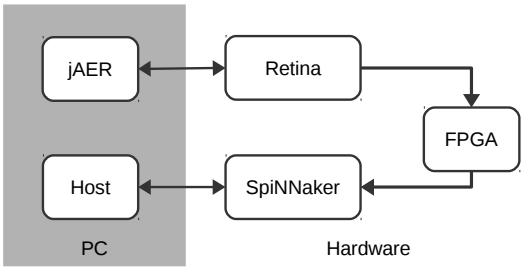
Although IT neurons are commonly described as narrowly selective object identifier, neurophysiological studies have shown a diverse selectivity of single neurons [16]. As illustrated in Figure 2(D), a single neuron (right) is modulated to both object identities and variables of identity-preserving transformations; if a population of such IT neurons tiles with the overlapping fashion (left), a more accurate recognition result containing the transformation parameter can be carried out with population coding. A simple weighted summation explains a wide range of invariant object recognition behaviour sufficiently [17], see Figure 2E.

## III. Preliminary Work

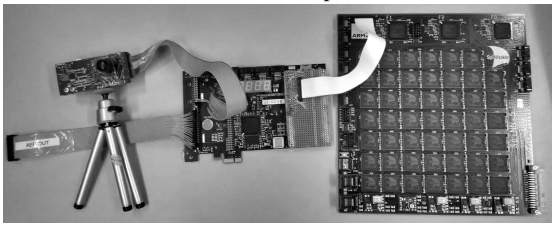
To explore how the brain may recognise objects in its general,accurate and energy-efficient manner, this preliminary work employs a neuromorphic hardware system formed from a DVS silicon retina in concert with the SpiNNaker real-time SNN simulator. Inspired by the behaviours of the primary visual cortex, Convolutional Neural Networks (CNNs) are modelled using both linear perceptrons and LIF neurons.

### A. Platform

The outline of the platform is illustrated in Figure 3a, where the hardware system is configured, controlled and monitored by the PC. The jAER [18] event-based processing software on the PC configures the retina and displays the output spikes through a USB link. The host communicates to the SpiNNaker board via Ethernet to set up its runtime parameters and to download the neural network model offline. It visualises [19] the spiking activity of the network in real-time. The photograph of the hardware platform, Figure 3b, shows that the silicon retina connects to the SpiNNaker 48-node system via a Spartan-6 FPGA board [20].




(a) Outline of the platform.



(b) Picture of the hardware platform.

Fig. 3: System overview of the object recognition platform.

### B. CNNs Models

There are two CNNs proposed to accomplish the dynamic hand posture recognition task. A straight forward method of template matching (V1-like neurons) is employed at first; The other multi-layer perceptrons (MLP) network is trained to improve the recognition performance.

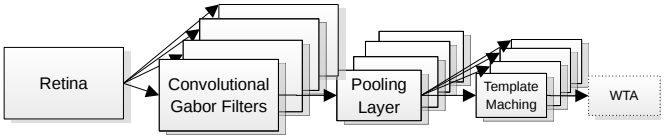



Fig. 4: Model 1.

**Model 1:** Template Matching. Shown in Figure 4 the first layer is the retina input, followed by the convolutional layer, where the kernels are Gabor filters responding to edges of four orientations. The third layer is the pooling layer where the size of the populations shrinks. This down-sampling enables robust classification due to its tolerance to variations in the precise shape of the input. The fourth layer is another convolution layer where the output from the pooling layer is convolved with the templates. The optional layer of Winner-Take-All (WTA) neurons enables a clearer classification result due to the inhibition between the neurons.

**Model 2:** Trained MLP. Inspired by the research of Lecun [21], we designed a combined network model with MLP and CNN (Figure 5). The first three layers are the same as the previous model. The training images for the 3-layered MLP are of same size and the posture is centred in the images. Therefore, a tracking layer is required to find the most active region and forward the centred image to the next layer. Attention effects were discovered in V4 by Moran and Desimone [22], which will be investigated in future work.

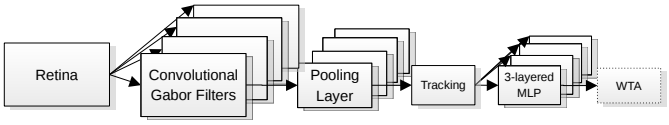



Fig. 5: Model 2.

### C. Moving from Perceptrons to Spiking Neurons

It remains a challenge to transform traditional artificial neural networks into spiking ones. There are attempts [23] [24] to estimate the output firing rate of the LIF neurons under certain conditions. The response function of the LIF neuron with Poisson input spike trains is given by the Siegert function [25], which is used in this work to estimate the synaptic weights.

Still there are some limitations on the response function. For the diffusion process, only small amplitude (weight) of the PostSynaptic Potentials (PSPs) generated by a large amount of input spikes (high spiking rate) work under this circumstance; plus, the delta function is required, i.e. the synaptic time constant is considered to be zero. Thus only a rough approximation of the output spike rate has been determined. Secondly, given different input spike rate to each pre-synaptic neurons, the parameters of the LIF neuron and the output spiking rate, how to tune every single corresponding synaptic weight remains a difficult task. Therefore, biologically plausible learning algorithms such as Spike Timing Dependent Plasticity (STDP) will be exploited in the future work.

### D. Experiments

**Experiment Set-up.** In order to evaluate the cost and performance trade-offs in optimizing the number of neural components, both the convolutional models described above are tested at different scales. Five videos of every posture are captured from the silicon retina in AER format, all of similar size and moving clock-wise in front of the retina.

Model 1 is tested on both perceptrons in Matlab and LIF neurons on SpiNNaker; whereas, Model 2 is only validated on Matlab since it merely estimates the highest performance of the system but in a non-biological way. All the details can be found in the latest submitted paper.

**Results.** In this study's largest configuration using these approaches, a network of 74,210 neurons and 15,216,512 synapses is created and operated in real-time using 290 SpiNNaker processor cores in parallel and with 93.0% accuracy. A smaller network using only 1/10th of the resources is also created, again operating in real-time, and it is able to recognise the postures with an accuracy of around 86.4% - only 6.6% lower than the much larger system. The recognition rate of the smaller network developed on this neuromorphic system is sufficient for a successful hand posture recognition system, and demonstrates a much improved cost to performance trade-off in its approach.

### E. Achievements

- Q. Liu and S. Furber, "Real-time recognition of dynamic hand posture on a neuromorphic system." Artificial Neural Networks ICANN 2015. (Under review)
- The 2014 CapoCaccia Cognitive Neuromorphic Engineering Workshop, Alghero, Sardinia, Italy
- From Maps to Circuits: Models and Mechanisms for Generating Neural Connections, Edinburgh UK

## IV. Future Work

The proposed research plan is illustrated in Figure 6, where the scope of the research is estimated in three dimensions. To build a invariant object recognition system using spiking neurons, this work will be completed in three stages:

- Year 1, building a position-invariant object recognition system exploiting V1-like neurons to classify five hand postures. (Completed)
- Year 1.5, combining scale- with position-invariance on the object recognition system, and building the hierarchy ventral pathway to the V2/V4 layer to recognise 50 simple combined features such as gratings and contours.
- Year 2.5, integrating position-, scale- and view-invariance by modelling the hierarchical visual pathway up to the IT cortex and equipping the system with the ability to recognise 200 objects in real time.

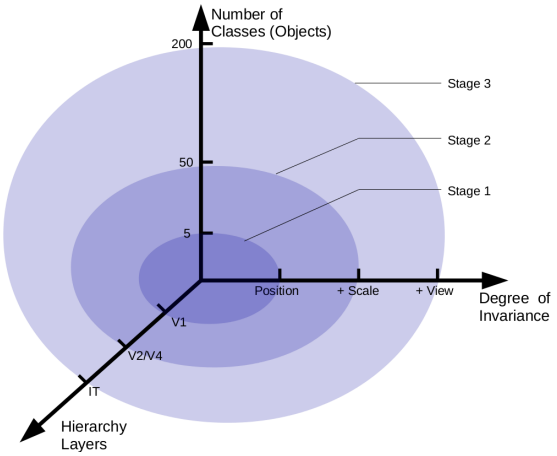



Fig. 6: 3D representation of the research plan on the transformation-invariant object recognition system. Three milestones are pointed out indicating the expected targets of the object recognition networks.

This work will contribute to the understanding of biological visual processing by means of mimicking the neural activities in the ventral stream. More importantly, the research will apply the accurate, rapid and robust approaches to artificial systems by exploring the brain's invariant object recognition. The performance of the real-time recognition system will be tested on each milestone to validate the success of the models. The neural activities and recognition rate will also be compared with biological data.

### A. Modelling the Ventral Visual Pathway

As the visual information propagates through the ventral stream (via visual area V1, V2/V4 and IT), neurons become selective for increasingly complex features. Along with this growing complexity of the preferred stimulus, neurons become more and more tolerant to the position and scale of the stimulus within their receptive fields. We will explore the invariant object recognition in three features: position, scale and viewing angle.

*1) Position Invariance:* Position invariance in the lower level of V1-liked neurons has been achieved in the preliminary work by convolving receptive fields with Gabor kernels. The following work in accordance with Figure 6 will focus on expanding the position invariance to higher hierarchical levels of the ventral stream.

*2) Scale Invariance:* Similar to orientation detection, V1 provides overcomplete population re-representations of visual image on the features of scale, frequency and orientation. It forms the basis of scale invariant object recognitions. Likewise, integrating the features into the higher abstraction of layered network to recognise more complex figures will require a tense work on tuning.

*3) View Invariance:* A difficult specificity-invariance trade-off occurs in view invariant recognition tasks, since the recogniser should be able to discriminate different objects while at the same time also tolerating to viewing angle transformations. Learning will play a very important role in this work, where objects observed with multiple view points can be recognised even if only single view point is presented during training.

### B. Size Scaling

The milestones set for the dimension of number of classes/objects is in accordance with experimental data from the study of neuroscience. In work by [26], the classical receptive field of the V2 cell consists of 48 grating stimuli and 80 contour stimuli; while Zoccolan et al. [27] tested and recorded the activity of the IT neurons of monkeys with 213 grayscale pictures of isolated real-world objects.

Thanks to the massively-parallel neural simulations possible in the SpiNNaker system, implementing real-time invariant object recognition becomes possible. However, it also requires the supporting software development to support larger neural networks than currently possible.

### C. Integration

To reach the milestone of building an object recognition system with position, scale and view invariance, integration of these separate models will be a challenge. It not only requires placing the models physically together but also merging their functions. As illustrated in Section II, single neurons are tuned to different features and object identities. This work requires further investigation into population coding and learning.

### D. Tuning

Tuning is the key to make the object recognition system a success. In preliminary work, Siegert transformation functions are used to adjust perceptral weights for spiking LIF neurons. This is a strong indicator of the feasibility of the work. However, learning algorithms such as STDP in spiking neural networks are must be employed to make the system more biologically plausible. It is hoped that, this work will provoke further study of learning algorithms on SpiNNaker.

### E. Benchmarking Performance

The performance of the real-time recognition system will be evaluated of each milestone to validate the success of the models. The neural activities and recognition rate will be compared with biological data which will act as a benchmark.

*1) Building a Dataset:* Building a well-labelled retinal output dataset is essential in spike-based object recognition study. Unified benchmarks with AER format will be ideal for SNN study, because of its non-frame, event-based fashion. These benchmark datasets will also make it possible for other researchers to test models without a silicon retina present.

*2) Testing/Comparing:* The testing and comparing on the dataset will verify the reliability of the models. The neural responses of single or populated neurons to the same dataset will be analysed in firing rate and response time. By comparing with the biological data, the model can be rectified.

## REFERENCES


[1] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?," *Neuron*, vol. 73, no. 3, pp. 415–434, 2012.

[2] M. Fabre-Thorpe, G. Richard, and S. J. Thorpe, "Rapid categorization of natural images by rhesus monkeys," *Neuroreport*, vol. 9, no. 2, pp. 303–308, 1998.

[3] C. Keysers, D.-K. Xiao, P. Földiák, and D. Perrett, "The speed of sight," *Journal of cognitive neuroscience*, vol. 13, no. 1, pp. 90–101, 2001.

[4] J. J. DiCarlo and D. D. Cox, "Untangling invariant object recognition," *Trends in cognitive sciences*, vol. 11, no. 8, pp. 333–341, 2007.

[5] J. A. Leñero-Bardallo, T. Serrano-Gotarredona, and B. Linares-Barranco, "A 3.6 s latency asynchronous frame-free event-driven dynamic-vision-sensor," *Solid-State Circuits, IEEE Journal of*, vol. 46, no. 6, pp. 1443–1455, 2011.

[6] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker Project," 2014.

[7] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology*, vol. 148, no. 3, p. 574, 1959.

[8] A. Anzai, X. Peng, and D. C. Van Essen, "Neurons in monkey visual area v2 encode combinations of orientations," *Nature neuroscience*, vol. 10, no. 10, pp. 1313–1321, 2007.

[9] Y. Daniel, M. Zarella, and G. Burkitt, "Whither the hypercolumn?," *The Journal of physiology*, vol. 587, no. 12, pp. 2791–2805, 2009.

[10] F. T. Qiu and R. Von Der Heydt, "Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules," *Neuron*, vol. 47, no. 1, pp. 155–166, 2005.

[11] P. Dean, "Effects of inferotemporal lesions on the behavior of monkeys.," *Psychological bulletin*, vol. 83, no. 1, p. 41, 1976.

[12] K. Tanaka, H.-a. Saito, Y. Fukada, and M. Moriya, "Coding visual images of objects in the inferotemporal cortex of the macaque monkey," *J Neurophysiol*, vol. 66, no. 1, pp. 170–189, 1991.

[13] C. G. Gross, "Single neuron studies of inferior temporal cortex," *Neuropsychologia*, vol. 46, no. 3, pp. 841–852, 2008.

[14] E. L. Schwartz, R. Desimone, T. D. Albright, and C. G. Gross, "Shape recognition and inferior temporal neurons," *Proceedings of the National Academy of Sciences*, vol. 80, no. 18, pp. 5776–5778, 1983.

[15] G. Sary, R. Vogels, and G. A. Orban, "Cue-invariant shape selectivity of macaque inferior temporal neurons," *Science*, vol. 260, no. 5110, pp. 995–997, 1993.

[16] R. Desimone, T. D. Albright, C. G. Gross, and C. Bruce, "Stimulus-selective properties of inferior temporal neurons in the macaque," *The Journal of Neuroscience*, vol. 4, no. 8, pp. 2051–2062, 1984.

[17] N. Majaj, H. Najib, E. Solomon, and J. DiCarlo, "A unified neuronal population code fully explains human object recognition," *Computational and Systems Neuroscience (COSYNE)*, 2012.

[18] T. Delbruck, "Frame-free dynamic digital vision," in *Proceedings of Intl. Symp. on Secure-Life Electronics, Advanced Electronics for Quality Life and Society*, pp. 21–26, 2008.

[19] C. Patterson, F. Galluppi, A. Rast, and S. Furber, "Visualising large-scale neural network models in real-time," in *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pp. 1–8, 2012.

[20] F. Galluppi, K. Brohan, S. Davidson, T. Serrano-Gotarredona, J.-A. P. Carrasco, B. Linares-Barranco, and S. Furber, "A real-time, event-driven neuromorphic system for goal-directed attentional selection," in *Neural Information Processing*, pp. 226–233, Springer, 2012.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[22] J. Moran and R. Desimone, "Selective attention gates visual processing in the extrastriate cortex," *Science*, vol. 229, no. 4715, pp. 782–784, 1985.

[23] G. La Camera, M. Giugliano, W. Senn, and S. Fusi, "The response of cortical neurons to in vivo-like input current: theory and experiment," *Biological cybernetics*, vol. 99, no. 4-5, pp. 279–301, 2008.

[24] A. N. Burkitt, "A review of the integrate-and-fire neuron model: I. homogeneous synaptic input," *Biological cybernetics*, vol. 95, no. 1, pp. 1–19, 2006.

[25] A. J. Siegert, "On the first passage time probability problem," *Physical Review*, vol. 81, no. 4, p. 617, 1951.

[26] J. Hegdé and D. C. Van Essen, "Temporal dynamics of shape analysis in macaque visual area v2," *Journal of neurophysiology*, vol. 92, no. 5, pp. 3030–3042, 2004.

[27] D. Zoccolan, M. Kouh, T. Poggio, and J. J. DiCarlo, "Trade-off between object selectivity and tolerance in monkey inferotemporal cortex," *The Journal of Neuroscience*, vol. 27, no. 45, pp. 12292–12307, 2007.