Toward a Moving Object Identification by Spiking Neurons

Abstract

This is a sort of position paper and no specific results are shown at this moment. We now are going to start an international collaborative research project on the issue of "Moving Object Identification." As one of many possibilities, our team exploits spiking neurons. In this paper, we study, first, an implementation already reported in public, and then we explore a couple of our own implementations aiming to be more biologically plausible.

1 INTRODUCTION

This paper presents a considerations on identifying moving objects – the topic which we plan as a collaborative project between the Izmir Institute of Technology and our University. We are going to approach to this issue from various aspects of machine learning techniques, and our group explores some possibilities of using spiking neurons.

The goal of the project is moving object identification in a scenery taken from a video stream. We see, nowadays, surveillance video cameras ubiquitously set in our city. Then extraction of only moving objects from a static background helps us tremendously reduce necessary memories. The same goes for traffic control by video. Or, movement detection from soccer video will help coaches analyse their strategies as good or bad in their football game results. Further, a system for recognizing lip-movements or sign language to translate them into texts in natural language will contribute to design a useful device to assist those who have a hearing problem. Probably, a military application is also one of them.

Thus, on one hand, the issue is very practical. Indeed, we have had many reports from a practical perspective. Just name a few: application to a traffic monitoring by Cheung et al. (2004); to a video surveillance by Heikkila et al. (2004); to a sign language or lip-movement recognition by Rivet et al. (2009); for an analysis of football scenario by Utsumi et al. (2002), and so on.

On the other hand, it is also interesting to approach this issue from a point of cognitive neuroscience, that is, how human, or animal, identifies moving objects with their visual cortex system. We all know a moving object is easier to be recognized than an object quietly located in background. For example, a satellite moving in its orbit dimly visible in the sky is easier to be recognized than a small star of the same brightness. Why? Our future goal is to pursue this kind of issue. So, the goal of this paper is from this aspect rather than to design very efficient and effective such system.

1.1 Background subtraction

To design a system of identifying moving objects, or detecting motion, we have to separate the moving foreground from the quiet background. For the purpose, we usually use a so called *background subtraction.*

McIvor (2000) gives us a good survey of this technique, in which he cites and paraphrases Heikkila et al. (1999). Assuming I_t and B_t is gray-scale value of pixel of current image and background, respectively, at time t, "*A pixel is marked as foreground if*

$$| I_t - B_t | > \tau$$

where τ is a predefined threshold." Heikkila et al.'s orignal paper reads *"As a result a binary image is formed where active pixels are labeled with 1 and non-active ones with 0."* Then *"It is necessary to update the background image frequently in order to guarantee reliable motion detection."*

McIvor paraphrases this as: *"The background update is*

$$B_{t+1} = \alpha I_t + (1 - \alpha) B_t,$$

where α is kept small to prevent artificial tails forming behind moving object. Two background corrections are applied:

1. *If a pixel is marked as foreground for more than m of the last M frames, then the background is updated as $B_{t+1} = I_t$.*

2. *If a pixel changes state from foreground to background frequently, it is masked out from inclusion in the foreground.*

The former compensates for such as a sudden illumination changes, while the latter compensates for a small motions like tree branches or leaves swinging in a blowing wind.

2 EXPERIMENT

At the onset of our survey for the project, we were inspired by the approach by Wu et al. (2008). We now take a look at it.

2.1 Wu's approach

The task is to detect moving object from a static background which is given as a stream of video frames represented by $M \times M$ gray-scale pixels. It is assumed that there will be no noises such as illumination changes or unimportant small changes like tree branches.

The network comprises of three layers. The input layer R is receptor array arranged in 2-D rectangle with its size being $M \times M$. The intermediate layer is made up of two 2-D rectangle arrays of spiking neurons $N1$ and $N2$. The size of both $N1$ and $N2$ is the same as R, also $M \times M$.

We now denote the location of pixel in the 2-D rectangle array with subscript xy. Each receptor R_{xy} sends an electric current $I_{xy}(t)$, depending on its corresponding gray-scale, to the intermediate layer.

The connections are described as follows. The input R_{xy} is linked to intermediate $N1_{xy}$ with two synapses: one is via *excitatory synapse without delay*, and the other via *inhibitory synapse with delay*. The input R_{xy} is also linked to intermediate $N2_{xy}$, in the same way but this time, one connection is via *excitatory synapse with delay*, and the other via *inhibitory synapse without delay*. The delay is constant and take the same value for all synapses with delay.

In each connection from input to intermediate, weight value of inhibitory synapse and excitatory synapse are adjusted so that if gray-scale of the input receptor do not change then both N1 and N2 are silent; if the gray-scale increases then N1 fires while N2 quiet; and if the gray-scale decreases N1 is quiet but N2 fires instead.

The both intermediate neurons $N1_{xy}$ and $N2_{xy}$ send spike trains to the output neuron O_{xy}. Output neuron fires if N1 or N2 fires. As time goes, the total of incoming spikes is calculated as

$$\rho_{xy}(t) = (1/T) \sum_{t}^{t+T} S_{xy}(t),$$

where $S_{xy}(t)$ is 1 when $N1_{xy}$ or $N2_{xy}$ fires at time t. Then, *"plotting $\rho_{xy}(t)$ as a grey image, white areas indicate neuron groups with high firing rate. Drawing the outside boundaries of firing neuron groups, boundaries of moving objects are extracted,"* as authors put it.

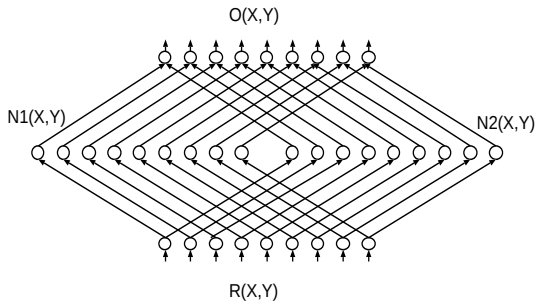

Figure 1: Wu's implementation. Video stream is given to input receptors via $M \times M$ grey scale pixels. Electric current according to the grey scale flows into intermediate spiking neurons $N1$ through excitatory synapse without delay and inhibitory synapse with delay, and to $N2$ through excitatory synapse with delay and inhibitory synapse without delay. Then the two neurons send spikes to the corresponding output spiking neuron. Note that all the straight lines from R_{xy} to both $N1_{xy}$ and $N2_{xy}$ actually represent two connections: excitatory and inhibitory.

2.2 Let's try in our way

As you may already noticed, the implementation is not a standard one. For example, Salinas et al. (2000) wrote: *"The output of a typical cortical neuron depends on the activity of a large number of synaptic inputs, several thousands of them, as estimated by anatomical techniques,"* citing Braitenberg et al. (1997), while in Wu's implementation, as mentioned above, each input receptor R_{xy} connects only to the two neurons $N1_{xy}$ and $N2_{xy}$ and then spike trains from these two neurons are given into only one output neuron O_{xy}.

We now check what we observe if we extract only one input receptor linked to two intermediate neurons and then fan-in to one output. Further we exploit here neural networks of *stochastic leaky integrate-and-fire neurons*, for some reason we will mention later. Then membrane potential $v_i(t)$ of neuron i at time t evolves in discrete time δt according to:

$$v_i(t) = v_i(t-\delta t)\exp(-\delta t/\tau_i) + \sum_j w_{ij}(t-\delta t) f_j(t-\delta t)$$

where τ_i is a time constant of neuron i, w_{ij} is synaptic weight value from neuron j to neuron i, and $f_j(t) = 1$ if neuron j fires at time t, otherwise 0.

As mentioned above, Wu et al. adjust synaptic weights manually. But it would not be easy, if not at all. According to Wu's implementation, one component from one input receptor via two intermediate neurons and then to one output neuron, two connections of which has axonal delay. As such, we have to specify 8 parameters, that is, 6 weights and 2 delays. Here we try a Genetic Algorithm to find an appropriate combination of these 8 parameters.

A population of chromosomes with 8 genes is to evolve with its fitness being as follows. Since in our experiment we have only one input receptor, if the input is from a background it keeps the same gray-scale value. When a part of moving object crosses the pixel a different gray-scale value is given to the receptor. Hence, output neuron should fire if and only if in the later case, otherwise should be quiet. So the number of coincidence of the actual time series of firing with the desired time series is the fitness. So far, however, we have not observed a good evolution.

2.3 To be more biologically plausible

Wu's implementation described above looks a little too artificial as authors put it *"Further research is required to establish the actual mechanisms employed by the visual cortex to determine motion."* Above all, each of the input pixel R_{xy} only connects to two neurons $N1_{xy}$ and $N2_{xy}$ and then spikes are given only to one output neuron O_{xy}. Network comprises with $M \times M$ such simple connections independently with each other. As already mentioned, this is not a standard implementation simulating biological visual system.

The network design we are currently experimenting is as follows. All of the input receptor R_{xy} connect to all intermediate spiking neurons of the same size N_{xy} with either of excitatory or inhibitory synapses all of which sent spikes to output spiking neurons with a delay. The strength of these synapses and delay are being gradually adjusted through learning, not specified manually at the beginning. In other words, the phenomena of background subtraction should emerge without a human design. This is toward an understanding how our brain identifies moving objects. We have already implemented some of its versions, but it has not so far been successful.

3 DISCUSSIONS

3.1 Is neural network intelligent?

What should be Artificial Intelligence?

McClelland (2009) wrote in his paper *"Even after more than a half a century of research on machine intelligence, humans remain far better than our strongest computing machines at a wide range of natural cognitive tasks, such as object recognition, ..."* Most of us agree, despite that we have had so many reports claiming an establishment of a machine intelligence.

At the same time, and more importantly to me, Frosini (2009) wrote *"Contradiction is often seen as a defect of intelligent systems and a dangerous limitation on efficiency."* Our behavior is not always optimally efficient. We sometimes make a mistake. Human intelligence is flexible and spontaneous, as Frosini tries to clarify *"the presence and importance of inconsistency in thought and in the processes trying to emulate it."*

If our goal is efficiency or effectiveness in identifying moving objects from video, *"kalman filtering is probably the most commonly used algorithm,"* as Heikkila (1999) wrote. See also Karmann et al. (1990). Or, Toyama et al. (1999) proposed what they call a *'Wallflower algorithm'* as a system for background maintenance, comparing it with eight other background subtraction algorithms, and claimed it outperforms all of the eights.

3.2 Options to link input to output

Assuming the task remained the same, that is, mapping input of R_{xy} onto the same size output O_{xy}, there might have many possible options of linking these two. What we should design is a system between these input and output. See Figure 2. Even if we limit our possibilities only to using spiking neurons, the white-box in the figure could be, such

as Random Recurrent Neural Network, Liquid Machine, Echo State Machine or else.

And then the question is which type of neurons are appropriate. For example, Izhikevich (2004) pointed out that each model has different types of dynamics, showing 20 such dynamics. In his experiment, Izhikevich (2006) specifically uses *regular spiking* type for excitatory neurons and *fast spiking* type for inhibitory neuron. What about other such combinations? Future work awaits to be explored.

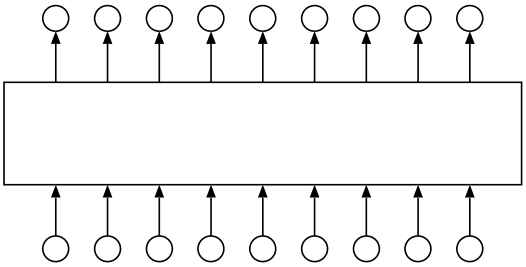

Figure 2: How we design a white-box between input image and output image?

4 FINAL REMARKS

The author know that even as a position paper, this is totally immature to submit to an international conference. However, since a series of these ongoing experiments is about to have more or less interesting results. So the author dare to decide to submit this paper for review. We expect stimulating discussions, assuming this position paper is generously accepted.

Reference

Braitenberg V., and A. Schüz (1997) "Cortex: statistics and geometry of neuronal connectivity." Springer.

Frosini, P. (2009) "Does intelligence imply contradiction?" Cognitive Systems Research, Vol. 10, No. 4. pp. 297–315.

Heikkila, M., M. Pietikäinen, and J. Heikkilä (2004) "A Texture-based Method for Detecting Moving Objects" Proceedings of British Machine Vision Conference, vol. 1, pp. 187–196.

Heikkila, J. and O. Silven (1999) "A real-time system for monitoring of cyclists and pedestrians." Proceedings of 2nd IEEE Workshop on Visual Surveillance, pp. 74–81.

Izhikevich, E. M. (2004) "Which model to use for cortical spiking neurons?" IEEE Transactions on Neural Networks, Vol. 15, pp. 1063–1070.

Izhikevich, E. M. (2006) "Polychronization: Computation with Spikes Export Find Similar." Neural Computation, Vol. 18, No. 2. pp. 245–282.

Karmann, K.-P., and A. von Brandt (1990) "Moving Object Recognition Using an Adaptive Background Memory." Time-varying Image Processing and Moving Object Recognition, No. 2, pp. 297–307.

McClelland, J. L. (2009) Is a Machine Realization of Truly Human-Like Intelligence Achievable? Cognitive Computation, Vol. 1, No. 1, pp. 17–21. Published online, Springer Science+Business Media

McIvor, A. M. (2000) "Background Subtraction Techniques." Proceedings of Image and Vision Computing pp. 147–153.

Salinas, E. and T. J. Sejnowski (2000) "Impact of Correlated Synaptic Input on Output Firing Rate and Variability in Simple Neuronal Models." The Journal of Neuroscience, Vol. 20, No. 16, pp. 6193–6209.

Sen-Chiong S., S.Cheung, and C. Kamath. (2004) "Robust Techniques for Background Subtraction in Urban Traffic Video." Journal: Visual Communications and Image Processing, pp. 881–892.

Soloyer, D., B. Rivet, L. Girin, C. Savariaux, J.-L. Schwartz, and C. Jutten, and S. D. Rivet (2009) "A study of lip movements during spontaneous dialog and its application to voice activity detection." Journal of Acoustic Society America. Vol. 125, No. 2, pp. 1184–1196.

Toyama, K., J. Krumm, B. Brumitt, and B. Meyers (1999) "Wallflower: Principles and Practice of Background Maintenance." Proceedings of 7th International Conference on Computer Vision, Vol. 1 (1999), pp. 255–261.

Utsumi, O., K. Miura, I. Ide. S. Sakai, and H. Tanaka (2002) "An Object Detection Method for Describing Soccer Games from Video." Proceedings of International Conference on Multimedia and Expo, Vol. 1, pp. 45–48.

Wu, Q., T. M. McGinnity, L. Maguire, and J. Cai (2008) "Motion Detection Using Spiking Neural

Network Model." Proceedings of the 4th international conference on Intelligent Computing: Advanced Intelligent Computing Theories and Applications - with Aspects of Artificial Intelligence. Lecture Notes In Artificial Intelligence (Springer) Vol. 5227. pp. 76–83.