Scotopic Visual Recognition

Bo Chen
California Institute of Technology
bchen@caltech.edu

Pietro Perona
California Institute of Technology
perona@caltech.edu

Abstract

Recognition from a small number of photons is important for biomedical imaging, security, astronomy and many other fields. We develop a framework that allows a machine to classify objects as quickly as possible, hence requiring as few photons as possible, while maintaining the error rate below an acceptable threshold. The framework also allows for a dynamic speed versus accuracy tradeoff. Given a generative model of the scene, the optimal tradeoff can be obtained from a self-recurrent deep neural network. The generative model may also be learned from the data. We find that MNIST classification performance from less than 1 photon per pixel is comparable to that obtained from images in normal lighting conditions. Classification on CIFAR10 requires 10 photon per pixel to stay within 1% the normal-light performance.

1. Introduction

Vision systems are optimized for speed and accuracy. Speed depends on the time it takes to capture an image (exposure time) and the time it takes to compute the answer. Computer vision researchers typically assume that there is plenty of light and a large number of photons may be collected very quickly[1]. Borrowing from the human vision literature we call this regime *photopic vision*. The image, while difficult to interpret, is (almost) noiseless; researchers ignore exposure time and focus on the trade-off between accuracy and computation time (e.g. see [6] Fig. 10).

Consider now the opposite situation, which we call *scotopic vision*[2], where photons are few and precious, and exposure time is long compared to computation time. The design tradeoff is between accuracy and exposure time [7], and computation time becomes a small additive constant.

Why worry about scotopic vision? We ask the opposite question: *“Why waste time collecting unnecessary photons?”* In some situations this question is compelling. First, one may be trying to sense/control dynamics that is faster than exposure time that guarantees good quality pictures, e.g. automobiles and quadcopters [5]. Second, in competitive scenarios, such as sports, a fraction of a second may make all the difference between defeat and victory [16]. Third, sometimes prolonged imaging has negative consequences, e.g. because phototoxicity and bleaching alter a biological sample [15] or because of health risks in medical imaging [9]. Fourth, sometimes there is little light in the environment, e.g. at night, and obtaining a good quality image takes a long time relative to achievable computational speed. Thus, we ask: “What is the minimal number of photons that are needed for good-enough vision?”, and “What is the best way to trade-off exposure time and accuracy?” and “How can one make visual decisions as soon as a sufficient number of photons has been collected?” It should be clear at this point that in scotopic vision photons are collected until the evidence is sufficient to make a decision.

While scotopic vision has been studied in the context of the physiology and technology of image sensing [1, 4], visual discrimination [8], and visual search [2], little is known regarding the computational principles for high-level visual tasks, such as categorization, in scotopic settings. Prior work on photon-limited image classification [18] deals with a single scotopic image, and does not study the trade-off between exposure time and accuracy. Moreover, scaling scotopic visual categorization even to modest-sized datasets, such as MNIST and CIFAR10 [11, 10], remains a challenge.

Our main contributions are:
1. A **computational framework** to study the trade-off between accuracy and response time in visual classification.
2. A **self-recurrent, deep convolutional architecture** to obtain **any-time**, quasi-optimal classification performance.
3. **Learning techniques** to train classifiers directly from a set of training data.

[1] In images with 8 bits per pixel of signal (i.e. SNR=256) pixels collect $10^4 - 10^5$ photons [14]. In full sunlight the exposure time is about 1/1000 s which is negligible compared to typical computation times.

[2] The term ‘scotopic vision’ literally means ‘vision in the dark’. It is usually associated to the physiological state where only rods, not cones, are active in the retina. We use this term to denote the general situation where a visual system is starved for photons, regardless the technology used to capture the image.

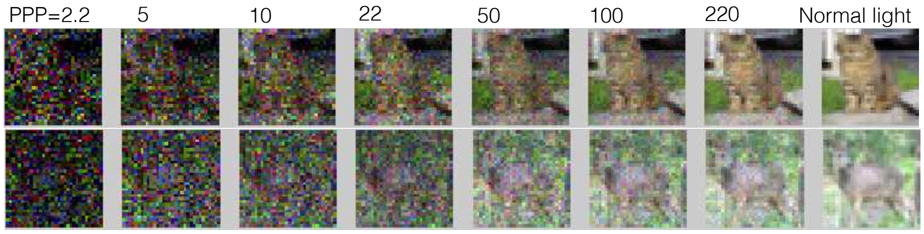

Figure 1. **Lowlight CIFAR10 images.** Sample synthetic lowlight images from the CIFAR10 dataset [10] with increasing average photons per pixel (PPP). PPP is proportional to the exposure time t.

2. Model

Our computational framework starts from a model of image capture. Each pixel in an image reports the brightness estimate of a cone of visual space by counting photons coming from that direction. The estimate becomes increasingly more accurate over time. Below, starting from a probabilistic assumption of the imaging process and of the target classification application, we describe a theory that allows for the best tradeoff between exposure time and classification accuracy.

We make three assumptions: 1) the world is stationary during the imaging process[3]; 2) photon arrival times follow a homogeneous Poisson process (details below); 3) a generative model of the task is available. We relax assumption (3) in Sec. 2.4.

Formally, the input $X(t) \in \mathbb{N}^d$ is an images with d pixels where $X_i(t)$ is the total number of photons arrived at pixel i in the time interval $[0, t]$. The task is to identify the corresponding visual category $Y \in \{0, 1, \dots, C\}$ of the image $X(t)$ with a given confidence level while minimizing exposure time t.

The pixels in the image are corrupted by several noise sources intrinsic to the camera [12]. We assume that the *fixed-pattern noise* and *quantization noise* are either negligible or removed by calibration. We focus on the *shot noise* and *dark current*. When the illuminance of the environment is fixed, photons arriving at each pixel i are assumed to follow a Poisson process with a constant rate. This rate is governed by both the true intensity of the stimulus mapping to the pixel and a small additive dark current. We assume that the dark current makes the darkest pixel emits photons at ϵ times the rate of the brightest pixel.

In addition, the average number of photons per pixel (PPP) is linear in t, and we will use time and PPP interchangeably. Since the information content in the image is directly related to the amount of photons, from now on we measure response time in terms of PPP instead of exposure time. Fig. 1 shows a series of images from the CIFAR10 dataset [10] with increasing PPP.

From Fig. 1 it is evident that images at different PPPs have different statistics. It would appear that a specialized system is required for each PPP level. Fortunately, one could exploit the structure of the input and build one system for images at all PPPs. The variation in the input $X(t)$ has two independent sources: one is the stochasticity in the photon arrival times, and the other the intra- and inter- class variation of the real intensity values of the object. Current computer vision techniques such as convolutional networks [11] excels at cases where only the first source of noise is present (i.e. when the photon counts has a high signal-to-noise ratio). The classical sequential probability ratio test (SPRT) developed by Wald [17] shows near-optimal speed-accuracy tradeoff when the second source of noise is absent.

We propose WaldNet, a deep network for speed-accuracy tradeoff (Fig. 2 (b-c)) that combines deep networks with SPRT. WaldNet assumes a generative model (Sec. 2.2) of the input $X(t)$ and uses it to compute the log posterior probability for each category (Sec. 2.3). This probability is fed into the SPRT (Sec. 2.1) to decide whether to collect more photons or terminate the task and report a decision. If a generative model is unavailable, WaldNet can be learned (Sec. 2.4) directly from low-light images.

2.1. Sequential probability ratio test

We first review SPRT [17]. Assuming a generative model of the images $X(t)$ is available, one can compute the log posterior probability of each visual category: $\log P(Y = c|X(t)), \forall c \in \{0, 1, \dots, C\}$. SPRT is a simple accumulation-to-threshold procedure as follows:

$$\begin{aligned} &\text{if } \log P(Y = c|X(t)) > \theta, \exists c : \text{report } Y = c \\ &\qquad\qquad\qquad \text{otherwise } : \text{increase } t \end{aligned} \tag{1}$$

When a decision is made, the declared class c has at least posterior probability e^θ according to the generative model, therefore the error rate of SPRT is at most $1 - e^\theta$.

For simple binary classification problems, SPRT is optimal in trading off speed versus accuracy in that no other algorithm can respond faster while achieving the same accuracy [17]. In the more realistic case where the categories are rich in intra-class variations, SPRT is shown to be asymptotic optimal, i.e. it gives optimal error rates as the exposure time becomes large [13]. Empirical study has also shown that even for short exposure times SPRT is near-optimal [3].

2.2. Deep generative model

SPRT requires full knowledge of a deep generative model, which is often impractical. However, we can spec-

[3] This is rather restrictive and will deserve a proper treatment in future studies.

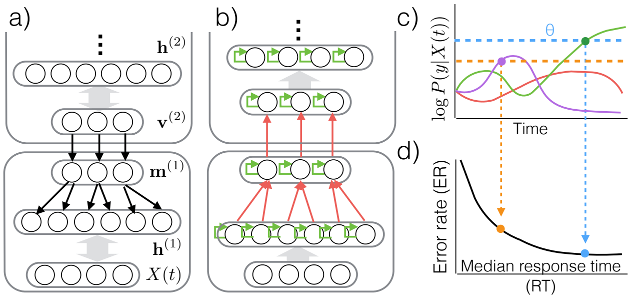

Figure 2. **WaldNet for lowlight visual recognition.** (a) A deep generative model (Eq. 2-4) of the sequence of input images of photon counts $X(t)$. Only the bottom stack and the lower part of the second stack are shown. Grey arrows represent full connections. Bi-directional arrows means full undirected connections. (b-c) WaldNet. (b) A deep network performs approximate inference on the generative model in (a) to compute the log posterior probability of the classes. Green arrows indicates accumulation of log likelihood ratios over time. (c) Sequential probability ratio test (SPRT) makes a decision as soon as the log posterior ratio crosses a threshold (two color-coded thresholds are shown in dash). (e) Speed accuracy tradeoff curves (sketch) produced by running WaldNet over many images and many repetitions of each image. Higher threshold leads to lower error.

ify a general architecture of the generative model and learn the parameters from data. The general structure we pick is a stack of smaller generative models. Each stack consists of an input vector $\mathbf{v}$, a hidden vector $\mathbf{h} \in \{0,1\}^{n_H}$ and a pooling vector $\mathbf{m} \in \{0,1\}^{n_M}$, and their connectivity is shown in Fig. 2(a).

We start from the bottom stack where the input is the cumulative raw photon counts $\mathbf{v} = X(t)$ over time t. Each binary element m_k of the pooling vector oversees a group G_k of hidden units. m_k represents the presence of an image feature within a spatial neighborhood G_k of the image, and h_j represents the exact location of the feature. Given the feature location h_j's, the input $X_i(t)$ at each pixel i for each point t in time is sampled according to

$$P(X_i(t)|\mathbf{h}) = Poiss(X_i(t)|\exp(\sum_j h_j W_{ij} + c_i)t) \quad (2)$$

where $W \in \mathbb{R}^{d \times n_H}$ and $c \in \mathbb{R}^d$ are weights and biases of the model.

The generative models at the higher layers of the stack is a deep belief network (DBN) [10] that models the activation probability of the max pooling units of the layer below. Details are in Appendix A.1.

2.3. Approximate inference

Given the input $X(t)$ at a specific time t, the generative model is a variant of the DBN, and the same techniques for efficient inference apply. Starting from the input $X(t)$, we can infer the hidden unit h_j of the bottom stack using:

$$S_{h_j}(t) \triangleq \log \frac{P(h_j = 1|X(t))}{P(h_j = 0|X(t))} = \sum_i W_{ij} X_i(t) + b_j t \quad (3)$$

where $b \in \mathbb{R}^{n_H}$ is the biases for the hidden units. In addition, we introduce a unit h_0 in each feature group to represent that the feature is absent. Naturally, $S_{h_0}(t) = 0$.

The log likelihood ratio of the pooling unit m_k is:

$$S_{m_k} \triangleq \log \frac{P(m_k = 1|X(t))}{P(m_k = 0|X(t))} \approx \max(0, \max_{j \in G_k}(S_{h_j}(t))) \quad (4)$$

which is equivalent to the standard max pooling and ReLu operations in modern deep networks.

S_{m_k} then becomes the input to the deep belief network above, which we use to infer the class label. Details of the inference procedure for DBNs are in Appendix A.1.

While Eq. 3 and Eq. 4 discuss the inference procedure where the input $X(t_0)$ and the model share the same exposure time t_0, in most scenarios, we would like to classify input at arbitrary time t_1 using models trained with images at a different time $t_0 > t_1$. Fortunately, this can be done by marginalizing out the unobserved data from t_1 to t_0. The marginalization (Appendix A.2) results in the following approximation for the hidden unit log probabilities:

$$S_{h_j} \approx \sum_i W_{ij} X_i(t_1) + b_j t_0 + a_j(t_1 - t_0) \quad (5)$$

where a_j is a scalar for each hidden unit h_j, and nontrivial to compute. Instead we learn a_j from data (see below).

In conclusion, the procedure to infer the class variable from images generated by a deep generative model resembles the computational flow in deep networks [10]. Thus, we can borrow the deep learning machinery for effective learning.

2.4. Learning

As discussed in Sec. 2.2, in practice, parameter learning is typically required for WaldNet with a predefined architecture (number of stacks, number of units in each stack, etc). We maximize the log posterior of the correct class $\log P(y = c|X(t))$ using stochastic gradient descent. We use images from multiple exposure times where the weight W is shared across time, and the bias b are estimated independently per time. Then we solve for the a_j's from the bias estimates using Eq. 5. See Appendix A.3.

3. Experiments

We study the effectiveness of WaldNet in trading off speed versus accuracy in two recognition tasks. We

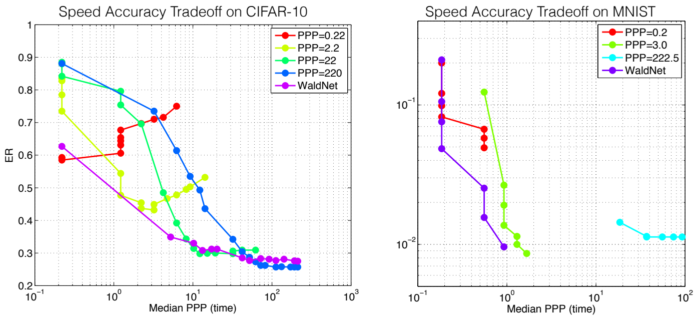

Figure 3. **Speed accuracy tradeoff of WaldNet**. CIFAR10 (left) and MNIST (right). WaldNet is trained on images from multiple PPP levels. Models labeled with PPP= x are "specialists" that have the same model complexity as WaldNet but trained using images from PPP=x only, and are used to access the optimal ER achievable by models of the same architecture at those PPPs. WaldNet makes decisions typically in less than 1 photon per pixel.

consider two standard datasets: MNIST [11] and CIFAR10 [10]. We synthesize lowlight images by generating a sequence of photons for each pixel, treating the pixel values as the ground truth intensity. We set the dark current $\epsilon = 3\%$. The brightest image we synthesize has about 2^8 photons, which corresponds to a pixel-wise maximum signal-to-noise ratio of 16 (4-bit accuracy), whereas the original MNIST images has (7 to 8-bit accuracy) that corresponds to 2^{14} to 2^{16} photons. We train WaldNet using images with light levels at PPP$\in \{0.2, 3.0, 220\}$ for MNIST, and $\{0.2, 2.2, 22, 220\}$ for CIFAR10, where the PPP is provided in tandem with the images so that network can adjust the biases according to Eq. 5. Details of model architecture and training protocol are in Appendix. A.3.

As a baseline, we train an ensemble of "specialist" models. Each specialist is a deep convolutional network with the same model architecture as the WaldNet, but is trained using only images at a single PPP. While these specialist models can not in theory be applied to images at different PPPs, we use a sensible strategy to do so. We scale their biases linearly with time to account for the increase in magnitude in the input (note that this is almost identical to Eq. 5 apart from using the wrong scaling factor). As the number of specialists approaches infinity, the ensemble gives a performance upper bound on WaldNet.

As shown in Fig. 3, WaldNet is very close to the best performance of the ensemble of specialists in trading off speed versus accuracy, despite the fact that the ensemble uses $3-4$ times the parameters. Overall, to stay with in 1% degradation from the optimal performance, WaldNet only requires about 10 PPP in CIFAR10 and < 1 PPP in MNIST.

4. Discussion and Conclusions

'Scotopic vision' is vision starved for photons. This happens when available light is low, and image capture time is longer than computation time. In this regime vision computations start as soon as the shutter is opened, and algorithms should be designed to process photons as soon as they hit the photoreceptors. To our knowledge, our study is the first to explore the exposure time versus accuracy trade-off of visual classification, which is essential in scotopic vision.

The proposed WaldNet provides an efficient approach to combining photon arrival events over time to form a coherent probabilistic interpretation, which allows the model to make a decision as soon as sufficient evidence has been collected. The proposed algorithm may be implemented by a deep feed-forward network that is very similar to a deep convolutional network. Despite the similarity of architectures, an experimental comparison of our adaptive network with the conventional kind shows large performance differences, both in terms of model parsimony and response time.

References

[1] H. Barlow. A method of determining the overall quantum efficiency of visual discriminations. *The Journal of physiology*, 160(1):155–168, 1962. 1

[2] B. Chen, V. Navalpakkam, and P. Perona. Predicting response time and error rate in visual search. In *Neural Information Processing Systems (NIPS)*, Granada, 2011. 1

[3] B. Chen and P. Perona. Towards an optimal decision strategy of visual search. *arXiv preprint arXiv:1411.1190*, 2014. 2

[4] T. Delbrück and C. Mead. Analog vlsi phototransduction. *Signal*, 10(3):10, 1994. 1

[5] E. D. Dickmanns. *Dynamic vision for perception and control of motion*. Springer Science & Business Media, 2007. 1

[6] P. Dollar, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *Submitted to IEEE Trans. on Pattern Anal. and Machine Intell.*, 2013. 1

[7] C. Ferree and G. Rand. Intensity of light and speed of vision: I. *Journal of Experimental Psychology*, 12(5):363, 1929. 1

[8] J. I. Gold and M. N. Shadlen. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36(2):299–308, Oct 2002. 1

[9] E. Hall and D. Brenner. Cancer risks from diagnostic radiology. *Cancer*, 81(965), 2014. 1

[10] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images, 2009. 1, 2, 3, 4

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1, 2, 4

[12] C. Liu, R. Szeliski, S. B. Kang, C. L. Zitnick, and W. T. Freeman. Automatic estimation and removal of noise from a single image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):299–314, 2008. 2

[13] G. Lorden. Nearly-optimal sequential tests for finitely many parameter values. *The Annals of Statistics*, pages 1–21, 1977. 2

[14] P. A. Morris, R. S. Aspden, J. E. Bell, R. W. Boyd, and M. J. Padgett. Imaging with a small number of photons. *Nature communications*, 6, 2015. 1

[15] D. J. Stephens and V. J. Allan. Light microscopy techniques for live cell imaging. *Science*, 300(5616):82–86, 2003. 1

[16] S. Thorpe, D. Fize, C. Marlot, et al. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996. 1

[17] A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945. 2

[18] M. N. Wernick and G. M. Morris. Image classification at low light levels. *JOSA A*, 3(12):2179–2187, 1986. 1