Chapter 4

SCOTOPIC VISUAL RECOGNITION

Sequential Reasoning without the Probabilistic Model

Our second project is scotopic visual recognition, which aims to recognize objects with as little light as possible. This project is motivated by real-world applications ranging from biological imaging to astrophysics. Unlike visual search (**Ch. 3**), most practical vision applications do not have the luxury of knowing the full probabilistic model for the task at hand. To circumvent this problem we proposed techniques to train a sequential algorithm directly to optimize the speed versus accuracy tradeoff (SAT).

4.1 Motivations

Just like biological systems, computer vision systems are optimized for accuracy and speed. Accuracy is well understood as the success rate at identifying object classes, estimating object poses, etc. 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, thus speed is limited by computation. This is called *photopic vision* where 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. Fig 10 of [1]).

> In images with eight bits per pixel of signal (i.e. SNR=256), pixels collect $10^4 - 10^5$ photons [2]. In full sunlight the exposure time is about 1/1000 s which is negligible compared to typical computation times.

Consider now the opposite situation, which we call *scotopic vision*, where photons are few and precious, and exposure time is long compared to computation time. As computation time becomes a small additive constant, the design tradeoff is between accuracy and exposure time [3]. There are multiple situations where trading off accuracy with exposure time is compelling. (1) One may be trying to sense/control dynamics that are faster than the exposure time that guarantees good quality pictures, e.g. automobiles and quadcopters [4]. (2) In competitive scenarios, such as sports, a fraction of a second may make all the difference between defeat

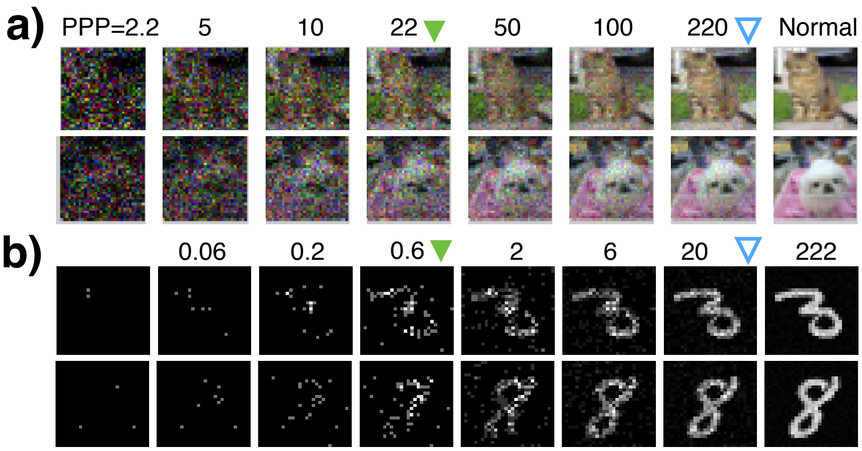

Figure 4.1: **Classification with few photons and speed-accuracy tradeoff.** Cumulative photon count N_t generated using sample images from the **(a)** CIFAR10 dataset and the **(b)** MNIST dataset with increasing average photons per pixel (PPP). PPP is proportional to the exposure time t. The images were obtained by simulating photon arrival times (**Sec. A.2**). Blue hollow arrows indicate the median PPP required for our scotopic classifier (WaldNet) to achieve comparable error rates (21%) as the model trained and tested using images under normal lighting conditions with about $2^7 \approx 10^4$ PPP (see **Sec. A.2** for protocol). Considerable speedups, of about two orders of magnitude, may be obtained by making classification happen as soon as a sufficient number of photons has been collected. Considerable further speed gains may be achieved by trading-off classification performance with decision speed: green solid arrows indicate the median PPP required to to maintain error rates below 22% for CIFAR and 1% for MNIST.

and victory [5]. (3) Sometimes prolonged imaging has negative consequences, e.g. because phototoxicity and bleaching alter a biological sample [6] or because of health risks in medical imaging [7]. (4) In sensor design, reduced photon counts allow for imaging with smaller pixels and ultra-high resolution [8], [9]. (5) 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, it is compelling to understand how many photons are needed for good-enough vision, and how one can make visual decisions as soon as a sufficient number of photons has been collected.

> The term 'scotopic / photopic vision' literally means 'vision in the dark / with plenty of light'. It is usually associated to the physiological state where only rods, not cones, are active in the retina. We use 'scotopic vision' to denote the general situation where a visual system is starved for photons, regardless of the technology used to capture the image.

Our work is further motivated by the recent development of *photon-counting imaging sensors*: single photon avalanche diode arrays [10], quanta image sensors [9], and gigavision cameras [8]. Instead of returning a high-quality image after a fixed exposure time, these sensors detect and report single photon arrival events at high frequencies. This ability to manipulate photon acquisition with fine granularity makes photon-counting sensors ideal for scotopic vision applications. Current computer vision technology has not yet taken advantage of these sensors.

4.2 Contributions

While scotopic vision has been studied in the context of the physiology and technology of image sensing [11], [12], as well as the physiology and psychophysics of visual discrimination [13] and visual search [14], little is known regarding the computational principles for high-level visual tasks, such as categorization and detection, in scotopic settings. Prior work on photon-limited image classification [15] deals with a single image, and does not study the trade-off between exposure time and accuracy. Instead, our work explore scotopic visual categorization on modern datasets such as MNIST and CIFAR10 [16], [17].

Sequential testing has appeared in the computer vision literature [18]–[20] in order to *shorten computation time*. These algorithms assume that all visual information ('the image') is present at the beginning of computation, thus focus on reducing computation time in photopic vision. By contrast, *our work aims to reduce capture time* and is based on the assumption that computation time is negligible when compared to image capture time. The similarity between the two lines of work is therefore only superficial.

Our main contributions are:

1. We present a **computational framework** for scotopic classification that dynamically decides the image exposure time for SAT.
2. When a probabilistic model of the classification task is given, we design a feedforward architecture yielding **any-time, quasi-optimal** scotopic classification.
3. When the probabilistic model is not available, we propose a **learning algorithm** to train the architecture for optimizing the SAT.
4. We conduct a **robustness analysis** with respect to sensor noise in current photon-counting sensor prototypes.

4.3 Framework for scotopic classification

Quantized sensory input

Our computational framework starts from a model of the sensory input. Each pixel in an image reports the brightness estimate of a cone of visual space by counting photons coming from that direction. The estimate improves over time.

To begin we consider a simpler version of the problem where the assumptions (**Ch. 2**) are met for SPRT. We assume that 1) the world is stationary during the imaging process (this may be justified as many photon-counting sensors sample the world at $> 1kHz$ [8], [9]); 2) photon arrival times follow a homogeneous Poisson process (details below) and 3) a probabilistic classifier based on photon counts is available. Assumption 3) may not be satisfied for practical object recognition classification problems, therefore we discuss how to do without this assumption in **Sec. 4.3**.

Poisson noise model

Sensors are corrupted by several intrinsic noise sources [21]. **Shot noise**: the number of photons incident on a pixel i in the t-th time interval, $X_{t,i}$, follows a Poisson distribution whose rate λ_i (Hz) depends on both the pixel intensity $I_i \in [0, 1]$ and a **dark current** ϵ_{dc}:

$$P(X_{t,i} = k) = Poisson(k|\lambda_i t\Delta) = Poisson(k|\lambda^{\phi}\frac{I_i + \epsilon_{dc}}{1 + \epsilon_{dc}}t\Delta), \tag{4.1}$$

where λ^{ϕ} is the illuminance (maximum photon count per pixel) per unit time [2], [8], [21], [22]. During readout, the photon count is additionally corrupted first by the amplifier's **read noise**, which is an additive Gaussian, then by the **fixed-pattern noise** which may be thought of as a multiplicative Gaussian noise [23]. As photon-counting sensors are designed to have low read noise and low fixed pattern noise[9], [10], [22], we focus on modeling the shot noise and dark current only. We will show (**Sec. 4.4**) that our models are robust against all four noise sources.

> According to the stationary assumption there is no need to model *motion-induced blur*. Additionally, for simplicity we do not model *charge bleeding and cross-talk* in colored images, and assume that they will be mitigated by the sensor community [24].

When the illuminance λ^{ϕ} of the environment is fixed, the amount of photons is roughly linear in the exposure time t (**Eq. 4.1**). Hence we use the number of photons per bright pixel (PPP) interchangeably with the exposure time t. i.e.:

$$PPP = \lambda^{\phi} t \Delta. \tag{4.2}$$

PPP= 1 means that a pixel with maximum intensity has collected 1 photon. Since the information content in the image is directly related to the number of photons, from now on we measure response time in terms of PPP instead of exposure time. **Fig. 4.1** shows a series of images from the CIFAR10 dataset [16] with increasing PPP.

Sequential probability ratio test for scotopic classification

Assume that a probabilistic model is available to interpret the sensory input given the class label – either provided by the application or learned from labeled data using techniques described in **Sec. 4.3** – we can apply SPRT to classify the photon streams. Since the classification task may contain multiple categories, the SPRT formulation **Eq. 2.3** needs to be extended to handle multiple hypothesis testing [25], [26].

Let $S_c(\boldsymbol{X}_{1:t}) \triangleq \log \frac{P(C=c|\boldsymbol{X}_{1:t})}{P(C\neq c|\boldsymbol{X}_{1:t})}$ denote the class posterior probability ratio of the visual category C for photon count input $\boldsymbol{X}_{1:t}, \forall c \in \{1, \ldots, K\}$, and let τ be an appropriately chosen threshold. SPRT conducts a simple accumulation-to-threshold procedure to estimate the category $\hat{C}$:

$$\begin{aligned} &\text{Compute } c^* = \arg\max_{c=1,\ldots,K} S_c(\boldsymbol{X}_{1:t}) \\ &\text{if } S_{c^*}(\boldsymbol{X}_{1:t}) > \tau : \text{ report } \hat{C} = c^* \\ &\text{otherwise } : \text{increase exposure time } t. \end{aligned} \tag{4.3}$$

Static versus dynamic exposure time models

In essence, SPRT decides when to respond dynamically, based on the stream of observations accumulated so far. As a result of the trial-by-trial variation of the signal, the response time also varies trial by trial. This regime is called “ **free-response**” (FR), in contrast to the “ **interrogation**” (INT) regime, typical of photopic vision, where a fixed-length observation is collected for each trial [27]. The observation length may be chosen according to a training set and fixed a priori. In both regimes, the length of observation should take into account the cost of errors, the cost of time, and the difficulty of the classification task.

Despite the striking similarity between the two regimes, SPRT (the FR regime) outperforms the INT regime, as we prove here for the case where the observations are i.i.d., and demonstrate empirically in **Sec. 4.4**.

Theorem 1 ***Free-response is asymptotically better than interrogation****.* *Assume that a probabilistic model is given to compute $S(\boldsymbol{X}_{1:t})$, and $\boldsymbol{X}_t$ is i.i.d. in time. Consider an FR algorithm that runs SPRT on $S(\boldsymbol{X}_{1:t})$ and let ϵ_{FR} and T_{FR} be its error rate and stochastic decision time. Also consider an INT algorithm with a fixed-length observation of t_{INT} that achieves an error of ϵ_{INT}. We have that the Bayes risk (**Eq. 2.1**) of the FR algorithm is less than or equal to that of the INT algorithm. In other words, as $\eta \to 0$:*

$$\mathbb{E}[T_{FR}] + \eta\epsilon_{FR} \leq t_{INT} + \eta\epsilon_{INT}.$$

Proof We prove the statement for binary classification with equal prior ($K = 2$, **Eq. 2.3**, the proof extends trivially to larger K). Consider all $\boldsymbol{X}_{1:t}$ generated from the positive class $C = 1$. Given an error rate requirement ϵ_{FR}, the FR algorithm sets up its threshold τ such that all the trials that terminate with $\hat{C} = 1$ must achieve a posterior probability of $1 - \epsilon_{FR}$, i.e. $P(C = 1|\boldsymbol{X}_{1:t}) = 1 - \epsilon_{FR}$, where $P(C = 1|\boldsymbol{X}_{1:t}) = Sigm(S(\boldsymbol{X}_{1:t}))$. Therefore, the threshold satisfies $Sigm(\tau) = 1 - \epsilon_{FR}$.

Since $\boldsymbol{X}_t$ is i.i.d. in time, $S(\boldsymbol{X}_{1:t}) = \sum_t S(X_t)$. Let $\mu \triangleq \mathbb{E}[X_t], \forall t$ represent the mean evidence accumulation rate (constant over time). The expected run time for the FR algorithm is

$$t_{FR} = \mathbb{E}[T_{FR}] = \frac{\tau}{\mu}.$$

Now consider an INT algorithm *with the same observation time* as the expected observation time for the FR algorithm, i.e. $t_{INT} = t_{FR}$. As $\eta \to 0$, $\epsilon_{FR} \to 0$, $t_{FR} \to \infty$ and $S(\boldsymbol{X}_{1:t_{FR}}) \geq 0$, a.s.. The error rate of the INT algorithm is

$$\begin{aligned} 1 - \epsilon_{INT} &= \mathbb{E}[Sigm(S(\boldsymbol{X}_{1:t_{FR}}))] \leq Sigm(\mathbb{E}[S(\boldsymbol{X}_{1:t_{FR}})]), a.s. \\ &= Sigm(\mu t_{FR}) = Sigm(\mu\frac{\tau}{\mu}) = Sigm(\tau) = 1 - \epsilon_{FR}, \end{aligned}$$

as a result of Jensen's inequality used on $Sigm(x)$, which is **concave** when $x \geq 0$.

Therefore as $\eta \to 0$, for any $t_{FR} = t_{INT}$, we have $\epsilon_{FR} \leq \epsilon_{INT}, a.s.$. Therefore for any pair of $\{t_{INT}, \epsilon_{INT}\}$ that minimizes Bayes risk for the INT algorithm, we can find an FR algorithm with $\{t_{FR}, \epsilon_{FR}\}$ that achieves a lower or equal Bayes risk. ∎

Computing class probabilities over time

The challenge of applying SPRT is to compute $S_c(\boldsymbol{X}_{1:t})$ for class c and the input stream $\boldsymbol{X}_{1:t}$ of variable exposure time t, or in a more information-relevant unit, variable PPP levels. Thanks to the Poisson noise model (**Eq. 4.1**), the sufficient statistics for observation $\boldsymbol{X}_{1:t}$ is the cumulative count $\boldsymbol{N}_t = \sum_{t'=1}^{t} \boldsymbol{X}_{t'}$ (visualized in **Fig. 4.1**), therefore we may rewrite $S_c(\boldsymbol{X}_{1:t})$ as $S_c(\boldsymbol{N}_t)$. It is evident that counts at different PPPs have different statistics. It would appear that a specialized system is required for each PPP level. This leads to the naive *ensemble* approach. Instead, we also propose a network called *WaldNet* that can process images at all PPPs and has the size of only a single specialized system. We describe the two approaches below.

> We insist on the need to distinguish between the cumulative count $\boldsymbol{N}_t$ and the conventional image, which is obtained by normalizing $\boldsymbol{N}_t$ to intensities within $[0, 255]$. By retaining the magnitude of the counts, $\boldsymbol{N}_t$ carries the uncertainty of the intensity estimates, which is crucial for evaluating the confidence of the class prediction.

A naïve approach: network ensembles

The simple idea is to build a separate model $S(\boldsymbol{N}_t)$ for the cumulative counts for each exposure time t (or light level PPP), either based on domain knowledge or learned from a training set. For best results one needs to select a list of representative light levels, and then apply each to input streams that were captured at the corresponding light level. For cumulative counts $\boldsymbol{N}_{t'}$ captured at light levels that are not on the list, one may simply apply the model with the closest light level. We refer to this as the 'ensemble' predictor.

One potential drawback of this ensemble approach is that training and storing multiple systems is *wasteful*. At different light levels, while the cumulative counts change drastically, the underlying statistical structure of the task stays the same. An approach that takes advantage of this relationship may lead to more parsimonious algorithms.

Model-based approach: WaldNet

An alternative is to exploit the knowledge about the cumulative counts across light levels. The variation in the input $\boldsymbol{N}_t$ 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. SPRT excels at reasoning about the first noise source while deep networks are ideal for capturing the second. Therefore we propose *WaldNet*, a deep network for speed-accuracy tradeoff (**Fig. 4.2b-c**) that combines deep networks with SPRT. Standard deep networks such as convolutional networks [17] (ConvNets) can not be applied directly as their inputs all have an identical exposure time T (e.g. $T \approx 33ms$ in normal lighting conditions). Instead, WaldNet utilizes lowlight noise statistics (**Sec. 4.4**) to adjust the computation within a deep network over exposure time t in order to compute the log class probability ratios $S_c(\boldsymbol{N}_t)$ over time t.

We first assume that a *generative* model for the cumulative counts $\boldsymbol{N}_T$ is available, and use it to develop a generative model for WaldNet. Then we provide a *discriminative* model with the identical computational form as the generative model, which may be learned directly from data.

> The generative model is rather technical. Readers who are not familiar with the literature on restricted Boltzmann machines and deep belief networks [28], [29] are encouraged to skip directly to the next section that discusses the discriminative training of WaldNet.

We assume that the generative model of input photon counts takes the form of a deep belief network [29]. The deep belief network is composed of multiple stacks. A stack on layer l consists of an input vector $\boldsymbol{v}^{(l)}$, a hidden vector $\boldsymbol{h}^{(l)} \in \{0, 1\}^{n_H^l}$ and a pooling vector $\boldsymbol{m}^{(l)} \in \{0, 1\}^{n_M^l}$. The log posterior ratio of the pooling vector of one layer becomes the input vector of the layer above, $v_i^{l+1} = \log \frac{P(m_i^{(l)}=1)}{P(m_i^{(l)}=0)}$, and the last pooling vector encodes desired log class posterior ratio $S(\boldsymbol{N}_T)$. $\boldsymbol{m}^{(l)}$, $\boldsymbol{h}^l$ and $\boldsymbol{v}^{(l)}$ are connected convolutionally as in a ConvNet, as follows:

1. Each pooling unit $m_k^{(l)}$ oversees a non-overlapping group $G_k^{(l)}$ of hidden units where at most one hidden unit is allowed to be on. $m_k^{(l)} = 1$ represents the presence of an image feature (say a 45° edge) anywhere within a spatial neighborhood $G_k^{(l)}$ of the image, and $h_j^{(l)} = 1$ indicates that the feature's location is j. This formulation is a generalization of probabilistic max pooling [30].

2. Each hidden unit $h_k^{(l)}$ connects to a small (say 5×5) neighborhood of input units $\boldsymbol{v}^{(l)}$. For layers $l > 1$ the hidden-input relationship is a standard RBM [28], [30], [31]. In the first layer where the input is the photon counts ($\boldsymbol{v}^{(1)} = \boldsymbol{N}_T$), the hidden-input relationship is a Poisson restricted Boltzmann machine [32], described

below. For notation simplicity we omit the layer superscript.

$$P(N_{i,T}|\boldsymbol{h}) = Poiss(N_{i,T}|\exp(\sum_j h_j W_{ij} + b_i^V)T), \tag{4.4}$$

where $W \in \mathbb{R}^{n_V \times n_H}$ and $b^V \in \mathbb{R}^{n_V}$ are weights and biases of the model. Since the connectivity is local, for each column in W, which corresponds to a hidden unit, only a small set (e.g. 25) of the entries are non-zero. The hidden units collectively model the mean firing rate $\lambda_i = \exp(\sum_j h_j W_{ij} + c_i^V)$ on location i.

Conversely conditioning on the cumulative photon count $\boldsymbol{N}_T$, the hidden units become independent and their distribution is given by:

$$P(h_j = 1|\boldsymbol{N}_T) = Sigm(\sum_i N_{i,T} W_{ij} + b_j^H). \tag{4.5}$$

Inference on the deep belief network faces one critical issue, which is that the observations are evolving over time, i.e. we need to compute $P(h_j = 1|\boldsymbol{N}_t)$ for any $t \leq T$, instead of merely the highly-exposed 'image' at time T. This may be done by marginalizing out the unobserved counts $\Delta \boldsymbol{N} \triangleq \sum_{t'=t+1}^{T} \boldsymbol{X}_{t'}$:

$$\begin{align} P(h_j = 1|\boldsymbol{N}_t) &= \sum_{\Delta \boldsymbol{N}} Sigm(\sum_i (N_{i,t} + \Delta N_i) W_{ij} + b_j^H) P(\Delta \boldsymbol{N}|\boldsymbol{N}_t) \tag{4.6} \\ &\approx Sigm(\sum_i (N_{i,t} + (T-t)\mathbb{E}[\lambda_i|N_{i,t}]) W_{ij} + b_j^H), \tag{4.7} \end{align}$$

where $\mathbb{E}[\lambda_i|N_{i,t}]$ is the estimated firing rate for location i. Using a Gamma prior $Gam(\mu_i t_0, t_0)$ on λ_i[1] we obtain that

$$P(h_j = 1|\boldsymbol{N}_t) \approx Sigm(\alpha(t) \sum_i W_{i,j} N_{i,t} + \beta_j(t)),$$

where $\alpha(t) \triangleq \frac{T+t_0}{t+t_0}$ and $\beta_j(t) \triangleq \frac{\tau(T-t)}{t+t_0} \sum_i W_{ij} \mu_i + b_j^H$ are two smooth scalar functions in t. Detailed derivations are in **Sec.** A.2.

Therefore, the log posterior ratio of the hidden units at the first layer is given by:

$$S_j^H(\boldsymbol{N}_t) \triangleq \log \frac{P(h_j = 1|\boldsymbol{N}_t)}{P(h_j = 0|\boldsymbol{N}_t)} \approx \alpha(t) \sum_i W_{i,j} N_{i,t} + \beta_j(t). \tag{4.8}$$

The log posterior ratio of the pooling unit m_k is:

$$S_k^M(\boldsymbol{N}_t) \triangleq \log \frac{P(m_k = 1|\boldsymbol{N}_t)}{P(m_k = 0|\boldsymbol{N}_t)} \quad = \mathcal{S}\max_{j \in G_k} \left(S_j^H(\boldsymbol{N}_t) \right) \approx \max_{j \in G_k} S_j^H(\boldsymbol{N}_t), \tag{4.9}$$

which is identical to the standard max pooling and the Maxout nonlinearity in deep networks [33], [34].

[1] We use a Gamma prior because it is the conjugate prior of the Poisson likelihood.

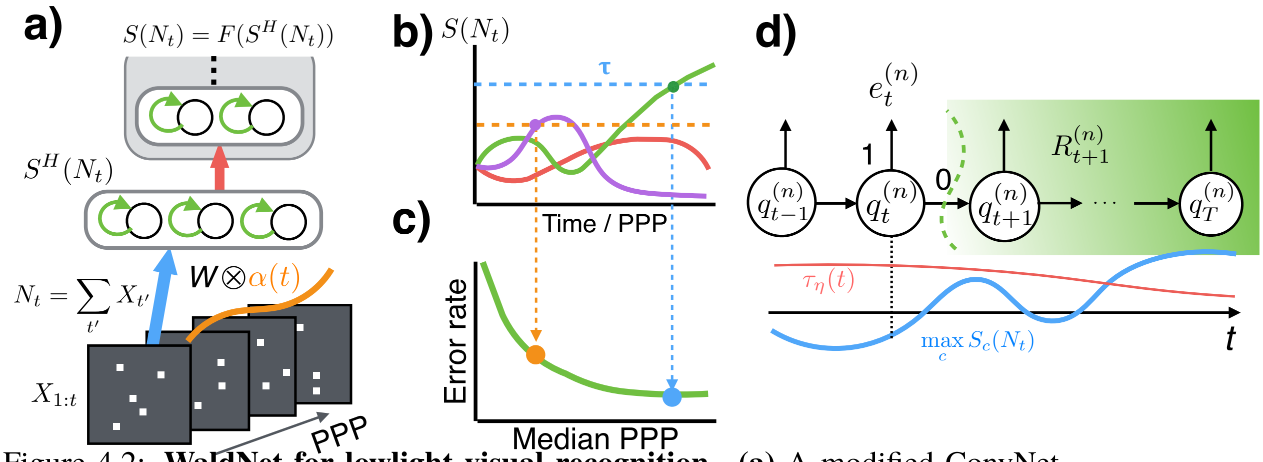

Figure 4.2: **WaldNet for lowlight visual recognition.** **(a)** A modified ConvNet for computing class posterior. The first layer is adapted (**Eq. 4.10**) to capture time-invariant features. From the cumulative photon counts $\boldsymbol{N}_t$ from duration $[0, t\Delta]$ (visualization in **Fig. 4.1**), WaldNet approximately computes hidden features $S^H(\boldsymbol{N}_t)$ that marginalize over unseen photons using weights $\boldsymbol{W}$ scaled by a time-varying scalar function $\alpha(t)$ (**Eq. 4.8**)). It then feeds the features into the remainder of the ConvNet F to compute log class posterior ratio $S(\boldsymbol{N}_t)$. **(b)** Deciding when to stop collecting photons. The class posteriors race to a common threshold to determine the category to predict. WaldNet stops photon collection as soon as one class crosses the threshold (**Eq. 4.3**). The example shows $S(\boldsymbol{N}_t)$ for three classes where the true class is green. Using a higher threshold (blue) yields a later but more accurate solution whereas a lower (orange) threshold is faster but risks misclassification. **(c)** The SAT curve (illustration only) produced by repeating (a-b) for multiple images and sweeping the threshold τ. **(d)** Learning time-varying threshold $\tau_\eta(t)$ (when class posterior learning (**Eq. 4.12**) is imperfect) to optimize Bayes risk with cost of error η (**Eq. 2.1**). The centipede network describes the recurrence relationship between risk $R_t^{(n)}$ starting from time t of example n and the risk $R_{t+1}^{(n)}$ starting from time $t+1$ (**Eq. 4.13**). $q_t^{(n)}$ is a gate (based on whether $S(\boldsymbol{N}_t)$ crosses threshold) that decides whether WaldNet stops at t with misclassification risk $e_t^{(n)}$ or continues collecting photons with risk $R_{t+1}^{(n)}$.

Discriminative training of WaldNet

Since the generative model may not be available in many practical applications, it may be more convenient to train a classifier that directly predicts the log posterior ratio $S(\boldsymbol{N}_t)$ and that shares the same computational structure as the inference procedure of the generative model. Fortunately the inference procedure bears striking similarity to a ConvNet, so that powerful deep learning tools (e.g. provided by the MatConvNet toolbox [35]) may be applied. Now we present the discriminative reasoning.

Inference procedure

Recall from the previous section that the inference procedure of WaldNet is an adjusted version of the standard ConvNet. In ConvNet, the input is an image $\boldsymbol{N}_T$ obtained from a fixed observation time T. ConvNet contains multiple layers of computations that may be viewed as a nesting of two transformations: (1) the first hidden layer $S^H(\boldsymbol{N}_T) = \mathbf{W}\boldsymbol{N}_T + \mathbf{b}^H$ that maps the input to a feature vector, and (2) the remaining layers $S(\boldsymbol{N}_T) = F(S^H_T)$ that map the features S^H to the log class posterior probabilities $S(\boldsymbol{N}_T)$. $\mathbf{W} \in \mathbb{R}^{D \times n_H}$ is a weight vector and $\mathbf{b}^H \in \mathbb{R}^{n_H}$ is a bias vector.

WaldNet differs from a ConvNet in two aspects. (1) The input $\boldsymbol{N}_t$ to a WaldNet is a *time-series* that includes the cumulative photon counts up to a moving horizon t, and the output $S(\boldsymbol{N}_t)$ is also a time-series, which encodes the log class posterior probabilities over time. (2) The first-layer features in WaldNet are computed differently *depending on the exposure time* t. The weights and biases of the transformation in S^H are adjusted smoothly over time using $\alpha(t) \in \mathbb{R}$ and $\boldsymbol{\beta}(t) \in \mathbb{R}^{n_H}$ (see **Eq. 4.8** and **Eq. 4.9**):

$$S^H(\boldsymbol{N}_t) = \alpha(t)\boldsymbol{W}\boldsymbol{N}_t + \boldsymbol{\beta}(t), \tag{4.10}$$

while the rest of the computations stays the same: $S(\boldsymbol{N}_t) = F(S^H(\boldsymbol{N}_t))$.

> The main intuition of our approach is that the stochasticity in photon arrivals is addressed with an exposure-time specific transformation S^H, and the intra- and inter- class variation is captured with an exposure-time invariant transformation F. The revised network has nearly the same number of parameters as a conventional ConvNet, but has the capacity to process inputs at different exposure times. The adaptation is critical for performance, as will be seen by comparison with simple rate-based methods in **Sec. 4.4**.

> Why do we single out the first layer features $S^H(\boldsymbol{N}_t)$ for adjustment? In theory features at any layer would do but it is more convenient at the first layer. This is because the adjustment procedure uses mean-field approximations and this (1) becomes increasingly less accurate as the feature computation becomes more nonlinear, and (2) requires computing the posterior mean of the feature, which may not have a handy closed form.

Training strategy

Recall that our goal is to train WaldNet to optimize the Bayes risk [36] (**Eq. 2.1**). In scotopic vision the Bayes risk R is formulated as

$$R \triangleq \mathbb{E}[t] + \eta\mathbb{E}[C \neq \hat{C}_t], \tag{4.11}$$

where $\mathbb{E}$[t] is the expected photon count required for classification, $\mathbb{E}[C \neq \hat{C}_t]$ is the error rate, and η describes the user's cost of error versus time. WaldNet asymptotically optimizes the Bayes risk provided that it can faithfully capture the log class posterior ratio $S(\boldsymbol{N}_t)$, and selects the correct threshold τ (**Eq. 2.3**). Sweeping η allows WaldNet to traverses the optimal SAT (**Fig. 4.2c**).

Our strategy is to separate training into two steps with distinct objectives: step one trains a WaldNet to approximate $S(\boldsymbol{N}_t)$, and step two picks the optimal threshold according to η to minimize the Bayes risk.

Step one: posterior learning
Given a lowlight dataset $\{\boldsymbol{N}_t^{(n)}, C^{(n)}\}_{n,t}$ where n indexes training examples and t indexes exposure time, we train the WaldNet to minimize:

$$-\sum_{n,t} \log P(C = C^{(n)} | \boldsymbol{N}_t^{(n)}, \mathcal{W}) + reg(\mathcal{W}), \tag{4.12}$$

where $\mathcal{W}$ collectively denote all the parameters in the WaldNet, and $reg(\mathcal{W})$ denotes $L2$ weight-decay on the filters. When a lowlight dataset is not available we simulate the dataset from intensity images according to the noise model in **Eq. 4.1**, where the exposure times are sampled uniformly on a logarithmic scale (see **Sec. 4.4**).

Step two: threshold tuning
After step one, if WaldNet captures the log class posterior ratios $S(\boldsymbol{N}_t)$, we can simply optimize a scalar threshold τ_η for each tradeoff parameter η. In practice, we may opt for a time-varying threshold $\tau_\eta(t)$ as step one may not be perfect.

> For instance, consider an adapted ConvNet that perfectly captures the class posterior. Ignoring the regularizer (right term of **Eq. 4.12**), we can scale up the weights and biases of the last layer (softmax) by an arbitrary amount without affecting the error rate, which scales the negative log likelihood (left term in **Eq. 4.12**) by a similar amount, leading to a better objective value. The magnitude of the weights are thus determined by the regularizer and may be off by a scaling factor. We therefore need to properly rescale the class posterior at every exposure time before comparing to a constant threshold, which is equivalent to using a time-varying threshold $\tau_\eta(t)$ on the raw predictions.

To learn the time-varying threshold $\tau_\eta(t)$, we need to formulate the Bayes risk objective as a function of $\tau_\eta(t)$. Let $\{\boldsymbol{N}_t^{(n)}\}_{t=1}^T$ be a sequence of lowlight images that are increasing in exposure time and generated from the n-th intensity image. Denote $q_t^{(n)} \triangleq \mathbb{I}[\max_c S_c(\boldsymbol{N}_t) > \tau_\eta(t)]$ the event that the posterior crosses decision threshold at time t, and $e_t^{(n)}$ the event that the class prediction at t is wrong. Let $R_t^{(n)}$ denote the Bayes risk of the sequence (indexed by n of the high-quality image $X^{(n)}$) incurred from time t onwards. $R_t^{(n)}$ may be computed recursively:

$$R_t^{(n)} = \Delta + \eta \left(q_t^{(n)} e_t^{(n)} + (1 - q_t^{(n)}) R_{t+1}^{(n)} \right), \tag{4.13}$$

where the first term is the cost of collecting photons during time interval $((t-1)\Delta t \Delta]$, the second term is the expected cost of committing to a decision that is wrong, and the last term is the expected cost of deferring the decision till more photons are collected.

The Bayes risk is obtained from averaging multiple photon count sequences, i.e. $R = \mathbb{E}[R_0^{(n)}]$. $q_t^{(n)}$ is non-differentiable with respect to the threshold $\tau_\eta(t)$, leading to difficulties in optimizing R. Instead, we approximate $q_t^{(n)}$ with a Sigmoid function,

$$q_t^{(n)}(\tau_\eta(t)) \approx Sigm\left(\frac{1}{\sigma_{temp}} (\max_c S_c(\boldsymbol{N}_t) - \tau_\eta(t)) \right), \tag{4.14}$$

where $Sigm(x) \triangleq 1/(1+\exp(-x))$, and anneal the temperature σ_{temp} of the Sigmoid over the course of training [37] (see **Sec. 4.4**).

> Even though we assume a certain form for the log class posterior ratio $S(\boldsymbol{X}_{1:t})$, this threshold learning procedure is very general and works for any $S(\boldsymbol{X}_{1:t})$. In particular, it may be used for learning SPRT procedures when the underlying probabilistic distribution is not i.i.d. in time.

4.4 Experiments

Exposure time versus signal

Our experiments use PPP interchangeably with exposure time t for performance measurement, since PPP directly relates to the number of bits of signal in each pixel (**Eq. 4.2**). In practice an application may be more concerned with exposure time. Thus it is helpful to relate exposure time, PPP and the bits of signal. **Table 4.1** describes this relationship for different illuminance levels. Derivations are in the Appendix **Sec. A.2**.

	Illuminance	exposure time t (s)					
Scene	E_v (LUX)	1/500	1/128	1/8	1	8	60
Moonless	10^{-3}					1.5	3
Full moon	1	0.5	1.5	3.5	5	6.5	8
Office	250	4.5	5.5	7.5	9	10.5	12
Overcast	10^3	5.5	6.5	8.5	10	11.5	13
Bright sun	10^5	9	10	12	13.5	15	16.5

Table 4.1: (Approximate) number of bits of signal per pixel under different illuminance levels. See Appendix for full derivation. For instance, in an office scene it takes 1/8 seconds to obtains a 7.5-bit image. Under full moon, the same high-quality image and the same sensor needs > 8 seconds to capture.

Baseline Models

We compare WaldNet against the following baselines:

Ensemble. We construct the ensemble (**Sec. 4.3**) using "specialist" models. Each specialist is a ConvNet with the same model dimensions (number of layers, number of hidden units of each layer, nonlinearity, etc) as the WaldNet, but is trained using only cumulative photon counts at a single PPP. We use four specialists with PPPs from $\{.22, 2.2, 22, 220\}$ respectively. To test cumulative counts $\boldsymbol{N}_{PPP'}$ with a PPP that is not on the training set, we rescale $\boldsymbol{N}_{PPP'}$ to have the same PPP as the specialist with the closest PPP. As the number of specialists grows, the ensemble approaches the best achievable SAT for WaldNet.

Photopic classifier. To justify the necessity of modeling photon count statistics in lowlight, we introduce another intuitive classifier. The classifier is a ConvNet trained on 'images' $\boldsymbol{N}_T$ from normal lighting conditions, and applied to properly rescaled cumulative counts $\boldsymbol{N}_t$ for $t \leq T$. We choose the specialist with PPP$= 220$ as the photopic classifier as it achieves the same accuracy as a network trained with 8-bit images.

Rate classifier. To test the significance of the uncertainty information carried by the cumulative counts, we train a classifier directly on the rate estimates without weight adaptation. Formally, the hidden unit on layer one is $S^H(\boldsymbol{N}_t) \approx \boldsymbol{W}\boldsymbol{N}_t/t + \boldsymbol{b}_j^H$. Note the similarity with our approximation used in **Eq. 4.8**.

We assume that all models have an internal clock, which enables the model to estimate the expected PPP under the constant illuminance assumption. When the illuminance changes, the model may rely on an independent external measure or the cumulative count itself to adjust PPP.

We consider two standard datasets: MNIST [17] and CIFAR10 [16]. We simulate

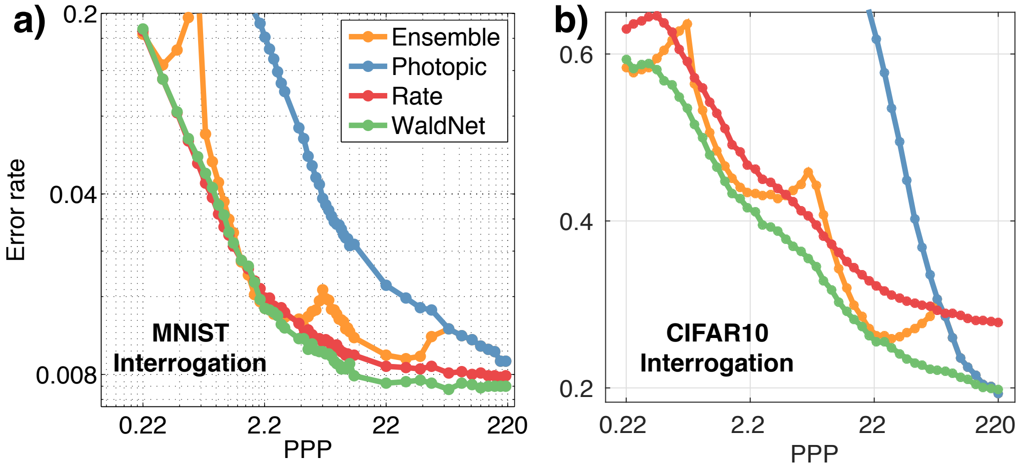

Figure 4.3: **Interrogation performance comparison**. Error rate plotted against the interrogation PPP for (**a**) MNIST and (**b**) CIFAR10. Each dot is computed from classifying $10k$ test examples with a fixed PPP.

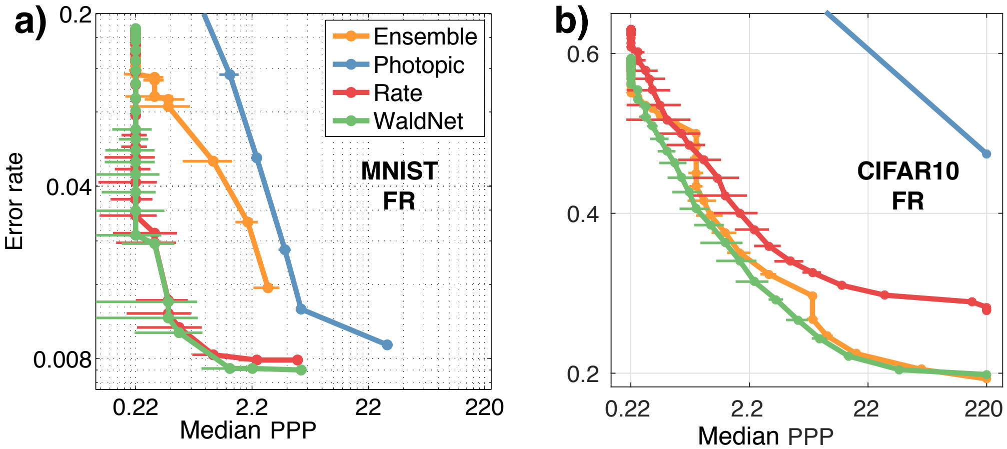

Figure 4.4: **Free response performance comparison**. Error rate plotted against *median* PPP for (**a**) MNIST and (**b**) CIFAR10. 1 bootstrap *ste* is shown for both the median PPP and error rate, the latter is too small to be visible.

lowlight image sequences using **Eq. 4.1**. MNIST contains gray-scaled 28 × 28 images of 10 hand-written digits. CIFAR10 contains 32 × 32 color images of 10 visual categories. The details of model architectures and training procedure are found in the Appendix **Sec. A.2**.

Results

The SAT curves in the INT regime are shown in **Fig. 4.3a** and b. Median PPP versus accuracy tradeoffs for all models in the FR regime are shown in **Fig. 4.4a** for MNIST and **Fig. 4.4b** for CIFAR10. All models use constant thresholds for

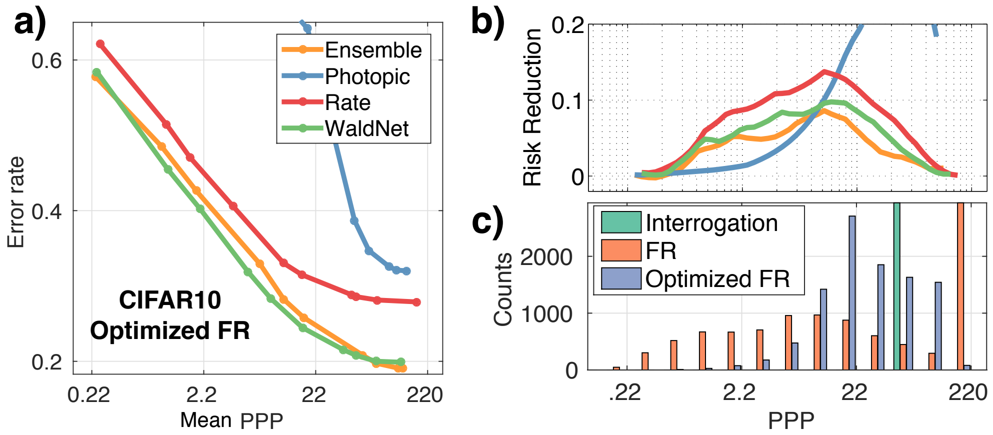

Figure 4.5: **Effect of threshold learning (Sec. 4.3)**. (**a**) Error rate against the *average* PPP for CIFAR10 using a network with optimized time-varying threshold $\tau_\eta(t)$. 1 bootstrapped *ste* is shown but not visible. (**b**) Each curve shows the Bayes risk reducation after optimization (**Sec. 4.3**, step two) per *average* PPP. (**c**) Response time (PPP) histograms under the interrogation, FR (before optimization), and FR (after optimization) of a WaldNet that achieves 22% error on CIFAR10.

producing the tradeoff curves. In **Fig. 4.5a** are average PPP versus accuracy curves when the models use the optimized dynamic thresholds (**Sec. 4.3**, step two).

Model comparisons

Overall, WaldNet performs well under lowlight. It only requires $<$ 1 PPP to stay within 0.1% (absolute) degradation in accuracy on MNIST and around 20 PPP to stay within 1% degradation on CIFAR10, even though recognition at such light levels (**Fig. 4.1**) may prove difficult for humans.

The ensemble was formed using specialists at logarithmically-spaced exposure times, thus its curve is discontinuous in the INT regime (**Fig. 4.3**). The peaks delineate transitions between specialists. The ensemble's performance at the specialized light levels $[.22, 2.2, 22, 220]$ also provides a proxy for the performance upper bound by ConvNets of the same architecture (apart from overfitting and convergence issues during learning). Using this proxy we see that even though WaldNet uses $1/4$ the parameters of the ensemble, it stays close to the performance upper bound. In FR regime, the ensemble is outperformed by WaldNet on MNIST (due to overfitting) and on par on CIFAR10 for lowlight conditions ($<$ 22 PPP). This showcases WaldNet's ability to handle photon counts at multiple PPPs without requiring explicit parameters (as it is the case for the ensemble).

The photopic classifier retrofitted to lowlight applications does not work well in

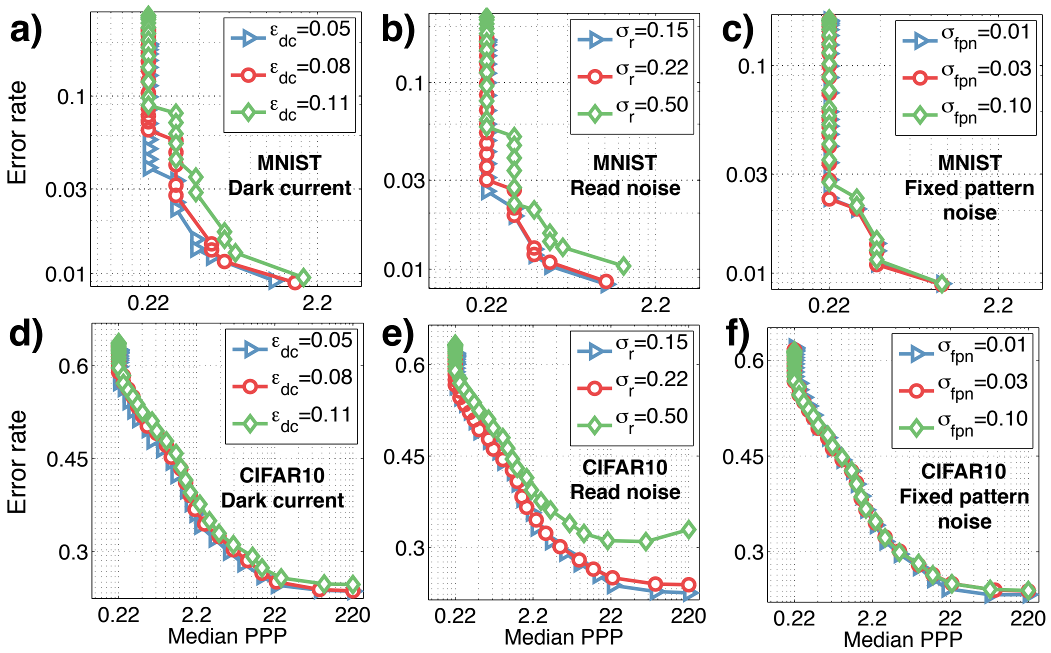

Figure 4.6: **Effect of sensor noise on WaldNet**. The rows correspond to datasets MNIST and CIFAR10, and the columns correspond to parameters of noise sources, which are the dark current ϵ_{dc}, the standard deviation of additive read noise σ_r, and the standard deviation of multiplicative fixed pattern noise σ_{fpn}. The baseline has $\epsilon_{dc} = 3\%$ and $\sigma_r = \sigma_{fpn} = 0$ for MNIST, and $\epsilon_{dc} = 5\%$, $\sigma_r = 0.22$ and $\sigma_{fpn} = 0.03$ for CIFAR10.

either dataset, which showcases the necessity of WaldNet as well as training with scotopic input. On MNIST, the photopic classifier also underperforms WaldNet in highlight regimes. This is because MNIST is rather easy to overfit, and training with lowlight inputs provides a form of regularization.

The rate classifier differs from WaldNet only in how the first layer feature is computed, thus the better performance of WaldNet in CIFAR10 is due solely to the WaldNet's time-adapted features (**Eq. 4.8**).

Effect of threshold learning

With constant thresholds (**Fig. 4.4**) WaldNet significantly outperforms the photopic classifier. As the latter has never seen any lowlight inputs, its assessment of the log posterior ratio is ill-suited to SPRT. Using learned dynamic thresholds (step two of **Sec. 4.3**) we see consistent improvement on the *average* PPP required for given error rate across all models (**Fig. 4.5b**), with more benefit for the photopic classifier. **Fig. 4.5c** examines the PPP histograms on CIFAR10 with constant (FR) versus dynamic threshold (optimized FR). We see with constant thresholds many decisions are made at the PPP cutoff of 220, so the median and the mean are vastly different. Learning dynamic thresholds reduce the variance of the PPP but make the median longer. This is ok because the Bayes risk objective (**Eq.** 2.1) concerns the average PPP, not the median. Clearly which threshold to use depends on whether the median or the mean is more important to the application.

Effect of INT versus FR

Cross referencing **Fig.** 4.3 and **Fig.** 4.4 reveals that FR with constant thresholds often brings 3x reduction in median photon counts. Dynamic thresholds also produce faster *average* and *median* responses. This is consistent with our theoretical result in **Theorem.** 1.

Sensitivity to sensor noise

Finally, we inspect how the network's performance is affected by sensor noise. For MNIST and CIFAR10, we take WaldNet and vary independently the dark current, the read noise and the fixed pattern noise (**Fig.** 4.6).

First, the effect of dark current and fixed pattern noise is minimal. Even an 11% dark current (i.e. photon emission rate of the darkest pixel is 10% of that of the brightest pixel) merely doubles the exposure time with little loss in accuracy. The multiplicative fixed pattern noise does not affect performance because WaldNet in general makes use of very few photons. Second, current industry standard of read noise ($\sigma_r = 22\%$ [9]) guarantees no performance loss. Lastly, the fact that $\sigma_r = 50\%$ hurts performance suggests that single-photon resolution is vital for scotopic vision (**Fig.** 4.6**b,e**).

4.5 Chapter summary

We proposed to study the important yet relatively unexplored problem of scotopic visual recognition. 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 should start as soon as the shutter is opened, and algorithms should be designed to process photons as soon as they hit the photoreceptors. While visual recognition from limited evidence has been studied [38], 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.

We proposed WaldNet, a model that combines photon arrival events over time to form a coherent probabilistic interpretation, and make a decision as soon as sufficient evidence has been collected. The proposed algorithm may be implemented by a deep feed-forward network similar to a convolutional network. Despite the similarity of architectures, we see clear advantages of approaches developed specifically for the scotopic environment. An experimental comparison between WaldNet and models of the conventional kind, such as photopic approaches retrofitted to lowlight images and ensemble-based approaches agnostic of lowlight image statistics, shows large performance differences, both in terms of model parsimony and response time (measured by the number of photons required for decision at desired accuracy). Finally, despite relying only on few photons for decisions, WaldNet is minimally affected by camera noises, making it an ideal model to be integrated with the recently-developed lowlight sensors.

References

[1] 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.

[2] 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*, vol. 6, 2015.

[3] C. Ferree and G. Rand, “Intensity of light and speed of vision: I.,” *Journal of Experimental Psychology*, vol. 12, no. 5, p. 363, 1929.

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

[5] S. Thorpe, D. Fize, C. Marlot, *et al.*, “Speed of processing in the human visual system,” *Nature*, vol. 381, no. 6582, pp. 520–522, 1996.

[6] D. J. Stephens and V. J. Allan, “Light microscopy techniques for live cell imaging,” *Science*, vol. 300, no. 5616, pp. 82–86, 2003.

[7] E. Hall and D. Brenner, “Cancer risks from diagnostic radiology,” *Cancer*, vol. 81, no. 965, 2014.

[8] L. Sbaiz, F. Yang, E. Charbon, S. Süsstrunk, and M. Vetterli, “The gigavision camera,” in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, IEEE, 2009, pp. 1093–1096.

[9] E. Fossum, “The quanta image sensor (qis): Concepts and challenges,” in *Imaging Systems and Applications*, Optical Society of America, 2011, JTuE1.

[10] F. Zappa, S. Tisa, A. Tosi, and S. Cova, “Principles and features of single-photon avalanche diode arrays,” *Sensors and Actuators A: Physical*, vol. 140, no. 1, pp. 103–112, 2007.

[11] H. Barlow, “A method of determining the overall quantum efficiency of visual discriminations,” *The Journal of Physiology*, vol. 160, no. 1, pp. 155–168, 1962.

[12] T. Delbrück and C. Mead, “Analog vlsi phototransduction,” *Signal*, vol. 10, no. 3, p. 10, 1994.

[13] J. I. Gold and M. N. Shadlen, “Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward,” *Neuron*, vol. 36, no. 2, pp. 299–308, Oct. 2002.

[14] B. Chen, V. Navalpakkam, and P. Perona, “Predicting response time and error rates in visual search,” in *Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 2699–2707.

[15] M. N. Wernick and G. M. Morris, “Image classification at low light levels,” *Journal of the Optical Society of America A*, vol. 3, no. 12, pp. 2179–2187, 1986.

[16] A. Krizhevsky and G. Hinton, *Learning multiple layers of features from tiny images*, 2009.

[17] 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.

[18] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, IEEE, vol. 1, 2001, pp. I–511.

[19] P. Moreels, M. Maire, and P. Perona, “Recognition by probabilistic hypothesis construction,” in *Computer Vision-ECCV 2004*, Springer, 2004, pp. 55–68.

[20] J. Matas and O. Chum, “Randomized ransac with sequential probability ratio test,” in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, IEEE, vol. 2, 2005, pp. 1727–1732.

[21] 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*, vol. 30, no. 2, pp. 299–314, 2008.

[22] E. R. Fossum, “Modeling the performance of single-bit and multi-bit quanta image sensors,” *Electron Devices Society, IEEE Journal of the*, vol. 1, no. 9, pp. 166–174, 2013.

[23] G. E. Healey and R. Kondepudy, “Radiometric ccd camera calibration and noise estimation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 16, no. 3, pp. 267–276, 1994.

[24] L. Anzagira and E. R. Fossum, "Color filter array patterns for small-pixel image sensors with substantial cross talk," *Journal of the Optical Society of America A*, vol. 32, no. 1, pp. 28–34, 2015.

[25] C. W. Baum and V. V. Veeravalli, "A sequential procedure for multihypothesis testing," *Information Theory, IEEE Transactions on*, vol. 40, no. 6, 1994.

[26] V. P. Dragalin, A. G. Tartakovsky, and V. V. Veeravalli, "Multihypothesis sequential probability ratio tests. ii. accurate asymptotic expansions for the expected sample size," *Information Theory, IEEE Transactions on*, vol. 46, no. 4, pp. 1366–1383, 2000.

[27] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen, "The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks.," *Psychological Review*, vol. 113, no. 4, p. 700, 2006.

[28] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, *et al.*, "Greedy layer-wise training of deep networks," *Advances in Neural Information Processing Systems*, vol. 19, p. 153, 2007.

[29] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[30] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, 2009, pp. 609–616.

[31] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[32] R. Salakhutdinov and G. Hinton, "Semantic hashing," *International Journal of Approximate Reasoning*, vol. 50, no. 7, pp. 969–978, 2009.

[33] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[34] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," *ArXiv preprint arXiv:1302.4389*, 2013.

[35] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, ACM, 2015, pp. 689–692.

[36] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[37] H. Mobahi and J. W. Fisher III, "On the link between gaussian homotopy continuation and convex envelopes," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, Springer, 2015, pp. 43–56.

[38] S. M. Crouzet, H. Kirchner, and S. J. Thorpe, “Fast saccades toward faces: Face detection in just 100 ms,” *Journal of Vision*, vol. 10, no. 4, p. 16, 2010.

[1] B. Chen and P. Perona, “Scotopic visual recognition,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2015, pp. 8–11.