Seeing into Darkness: Scotopic Visual Recognition

Bo Chen Pietro Perona
California Institute of Technology
bchen3, perona@caltech.edu

Abstract

Images are formed by counting how many photons traveling from a given set of directions hit an image sensor during a given time interval. When photons are few and far in between, the concept of 'image' breaks down and it is best to consider directly the flow of photons. Computer vision in this regime, which we call 'scotopic', is radically different from the classical image-based paradigm in that visual computations (classification, control, search) have to take place while the stream of photons is captured and decisions may be taken as soon as enough information is available. The scotopic regime is important for biomedical imaging, security, astronomy and many other fields. Here we develop a framework that allows a machine to classify objects with as few photons as possible, while maintaining the error rate below an acceptable threshold. A dynamic and asymptotically optimal speed-accuracy tradeoff is a key feature of this framework. We propose and study an algorithm to optimize the tradeoff of a convolutional network directly from low-light images and evaluate on simulated images from standard datasets. Surprisingly, scotopic systems can achieve comparable classification performance as traditional vision systems while using less than 0.1% *of the photons in a conventional image. In addition, we demonstrate that our algorithms work even when the illuminance of the environment is unknown and varying. Last, we outline a spiking neural network coupled with photon-counting sensors as a power-efficient hardware realization of scotopic algorithms.*

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 typically operates in the 'photopic' paradigm where the environment is well lit and an image may be acquired so rapidly that exposure time becomes negligible compared to computational times [1]. However, when the visual system is starved for photons (e.g. at night), exposure time could dominate the computational time [2]. We refer to this less-studied situation as 'scotopic' (**Fig. 1a**)[3]. Instead of waiting for high-quality imagery, scotopic visual systems should analyze photons as they stream in, and make decisions as soon as a sufficient number of photons has been collected (**Fig. 1a**).

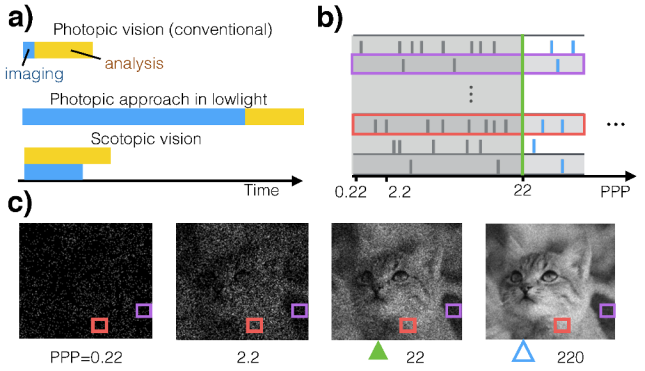

Figure 1. **Scotopic visual classification.** **a)** Computation time breakdown of photopic vs scotopic approaches. In conventional photopic approaches, image formation time ($\approx$ 30ms) is dwarfed by computation time. The same approach in twilight ($100\times$ darker) slows down sustaintially due to prolonged imaging time. The proposed scotopic approach reduces runtime by 1) analyzing input as photons stream in and 2) terminating photon collection as soon as sufficient information has been collected for the particular input. **b)** A sample photon stream. Each row coresponds to one pixel and each vertical bar is a photon arrival event (color-coded rows correspond to marked pixels in c)). The 'amount of light' that has been collected is quantified by the average photons per pixel (PPP), which is proportional to the exposure time t assuming constant illuminance. **c)** Cumulative photon counts at selective PPPs visualized as images. Blue hollow arrow indicates a typical median PPP required for our scotopic classifier (WaldNet, **Sec. 3.3.2**) to achieve a comparable error rate as the model trained and tested using images under normal lighting conditions with about $2^7 \approx 10^4$ PPP (see **Sec. 4.2** for protocol). Green solid arrow (and bar in b)) indicates the median PPP to stay within 1% performance degradation.

[1] e.g. In full sunlight an image with 8 bits per pixel of signal (or $10^4 - 10^5$ photons per pixel [32]) is collected in $1ms$.

[2] Illuminance at full moon is 10^{-5} of that in full sunlight, resulting in a 100 sec exposure time for an 8-bit image.

[3] We take the literal meaning of 'scotopic / photopic vision' ('vision in the dark / with plenty of light') as opposed to their biological definitions associated to the physiological state of whether cones are active in the retina.

Scotopic vision studies the tradeoff between accuracy and exposure time. It is compelling in situations such as 1) autonomous driving [11] and competitive robotics [41], where the desired response time does not guarantee good quality pictures, and 2) medical imaging / classification [38] and astrophysics [27], where photons are expensive due to phototoxicity or prolonged acquisition times.

Scotopic vision also gains prominence thanks to the recent development of *photon-counting imaging sensors*: single photon avalanche diode arrays [46], quanta image sensors [12], and gigavision cameras [36]. These sensors detect and report *single photon arrival events* in quick succession, an ability that provides fine-grained control over photon acquisition that is ideal for scotopic vision applications. By contrast, conventional cameras, which are designed to return a high-quality image after a fixed exposure time, produce an insurmountable amount of noise when forced to read out images rapidly and are suboptimal at low light. Current computer vision technology has not yet taken advantage of photon-counting sensors since they are still under development. Fortunately, realistic noise models of the sensors [12] are already available, making it possible (and wise) to innovate computational models that leverage and facilitate the sensor development. The challenge facing these models is to be compatible with the high sampling frequencies and the particular noises of the photon-counting sensors.

While scotopic vision has been studied in the context of the physiology and technology of image sensing [3, 10], as well as the physiology and psychophysics of visual discrimination [15] and visual search [5], 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 [45] deals with a single image, whereas our work not only studies the tradeoff between exposure time and accuracy, but also explores scotopic visual categorization on datasets of modern complexity.

Our main contributions are:
1. A **discriminative framework** for scotopic classification that can trade-off accuracy and response time.
2. A feedforward architecture yielding **any-time, quasi-optimal** scotopic classification.
3. A **learning algorithm** optimizing the speed accuracy tradeoff of lowlight classifiers.
4. **Robustness analysis** regarding sensor noise in current photon-counting sensors.
5. A **spiking implementation** that trades off accuracy with computation / power.
6. A **light-level estimation** capacity that allows the implementation to function without an external clock and at situations with unknown illuminance levels.

2. Previous Work

Our approach to scotopic visual classification of collecting only enough observations to make a decision descends from Wald's Sequential Probablistic Ratio Test (SPRT) [44]. Wald proved optimality of SPRT under fairly stringent conditions (see **Sec. 3**). Lorden, Tartakowski and collaborators [26, 40] later showed that SPRT is *quasi-optimal* in more general conditions, such as the competition of multiple one-sided tests, which turns out to be useful in multiclass visual classification.

Convolutional neural networks (ConvNets) [14, 22, 21, 19, 39] have achieved great success in image recognition problems. We show that vanilla ConvNets are inadequate for scotopic vision. However, they are very appropriate once opportune modifications are applied. In particular, our scotopic algorithm marries ConvNet's ability to classify high-quality images with SPRT's ability to trade off acquisition time with accuracy in a near-optimal fashion.

Sequential testing has appeared in the computer vision literature [43, 31, 28] 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. In addition, these algorithms either require an computationally intensive numerical optimization [33] or fail to offer optimality guarantees [47, 8]. In comparison, our proposed strategy has a closed-form and is asymptotically optimal in theory.

Recurrent neural networks (RNN) [18, 16] is a powerful tool for sequential reasoning. Our work is inherently recurrent as every incoming photon prompts our system to update its decision. However, conventional RNNs are inefficient at handling the sheer amount of data produced by photo-counting sensors (imaging at $1kHz$). Therefore, we employ a continuous-time RNN [23] that can be trained using samples at arbitrary times and show that a logarithmic number of (4) samples per photon stream suffice in practice.

VLSI designers have produced circuits that can signal pixel 'events' asynchronously [9, 10, 25] as soon as a sufficient amount of signal is present. This is ideal for our work since conventional cameras acquire images synchronously (all pixels are shuttered and A-D converted at once) and are therefore ill-adapted to scotopic vision algorithms. The idea of event-based computing has been extended to visual computation by O'Connor et al. [34] who developed an event-based deep belief network that can classify handwritten digits. The classification algorithms and the spiking implementation that we propose are distantly related to this work. Our emphasis is to study the best strategies to minimize response time, while their emphasis is on spike-based computation.

The pioneering work of [7] establishes a generative

framework for studying scotopic classification. By comparison we employ a discriminative framework that does not require a full probabilistic model of images. This gives us the flexibility to incorporate image preprocessing (for better classification accuracy), light-level estimation (for handling unknown and variable light levels) and out-of-distribution testing (for measuring robustness against realistic sensory noises). We additionally provide a power-efficient spiking network implementation for integration with photon-counting sensors.

3. A Framework for Scotopic Classification

3.1. Image Capture

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 improves over time. Starting from a probabilistic assumption of the imaging process and of the target classification application, we derive an approach that allows for the best tradeoff between exposure time and classification accuracy.

We make three assumptions (relaxed or tested later):
1. The world is **stationary** during the imaging process. This may be justified as many photon-counting sensors sample the world at $> 1kHz$ [36, 12]. Later we test the model under camera movements and show robust performance.
2. Photon arrival times follow a **homogeneous Poisson process**. This assumption is only used to develop the model. We will evaluate the model in **Sec. 4.4** using observations from realistic noise sources.
3. A **discriminative** classifier based on photon streams is available. We will discuss how to obtain such a model in **Sec. 3.4**.

Formally, the input $\mathbf{X}_{1:t}$ is a stream of photons incident on the sensors during time $[0, t\Delta)$, where time has been discretized into bins of length Δ. $X_{t,i}$ is the number of photons arrived at pixel i in the tth time interval, i.e. $[(t-1)\Delta, t\Delta)$. The task of a scotopic visual recognition system is two fold: 1) computing the category $C \in \{0, 1, \ldots, K\}$ of the underlying object, and 2) crucially, determining and minimizing the exposure time t at which the observations are deemed sufficient.

3.1.1 Noise Sources

The pixels in the image are corrupted by several noise sources intrinsic to the camera [24]. **Shot noise**: The number of photons incident on a pixel i in the unit time follows a Poisson distribution whose rate (in Hz) depends on both the pixel intensity $I_i \in [0, 1]$ and a **dark current** ϵ_{dc}:

$$P(X_{t,i} = k) = Poisson(k|\lambda_\phi \frac{I_i + \epsilon_{dc}}{1 + \epsilon_{dc}} t\Delta) \quad (1)$$

where the illuminance λ_ϕ is the expected photon count per bright pixel (intensity 1) per unit time [32, 36, 24, 13]. 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 [17]. As photon-counting sensors are designed to have low read noise and low fixed pattern noise[12, 46, 13], 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. Additionally, according to the stationary assumption there is no *motion-induced blur*. 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 [2].

A natural quantifier of the information content within a photon stream $\mathbf{X}_{1:t}$ is the average number of photons per bright pixel (PPP). PPP of $\mathbf{X}_{1:t}$ is estimated by dividing the average photon counts of pixels with true intensity 1 by the average scene illuminance λ_ϕ over duration $[0, t\Delta]$. **Fig. 1** shows a series of images with increasing PPP.

If the scene illuminance λ_ϕ is constant over time (which we assume to be true prior to **Sec. 3.5**), PPP is linear in the exposure time t:

$$PPP = \lambda^\phi t\Delta \quad (2)$$

hence we use exposure time t and PPP interchangeably. When scene illuminance fluctuates over time, an effective exposure time $\hat{t}$ may be estimated (see **Sec. 3.5**) based on a nominal illuminance so that the problem reduces to the case with constant illuminance.

3.2. Sequential probability ratio test

Our decision strategy for trading off accuracy and speed is based on SPRT, for its simplicity and attractive optimality guarantees. Assume that a probabilistic model is available to predict the class label C given a sensory input $\mathbf{X}_{1:t}$ of any duration t [4], SPRT conducts an accumulation-to-threshold procedure to estimate the category $\hat{C}$:

Let $S_c(\mathbf{X}_{1:t}) \triangleq \log \frac{P(C=c|\mathbf{X}_{1:t})}{P(C\neq c|\mathbf{X}_{1:t})}$ denote the class posterior probability ratio of the visual category C for photon count input $\mathbf{X}_{1:t}$, $\forall c \in \{1, \ldots, K\}$, and let τ be an appropriately chosen threshold. SPRT repeats the following:

$$\begin{aligned} &\text{Compute } c^* = \arg\max_{c=1,\ldots,K} S_c(\mathbf{X}_{1:t}) \\ &\text{if } S_{c^*}(\mathbf{X}_{1:t}) > \tau : \text{ report } \hat{C} = c^* \\ &\text{otherwise } : \text{increase exposure time } t. \end{aligned} \quad (3)$$

When a decision is made, the declared class $\hat{C}$ has a probability that is at least $\exp(\tau)$ times bigger than the probability of all the other classes combined, therefore the error rate

[4] either provided by the application or learned from labeled data using techniques described in **Sec. 3.4**

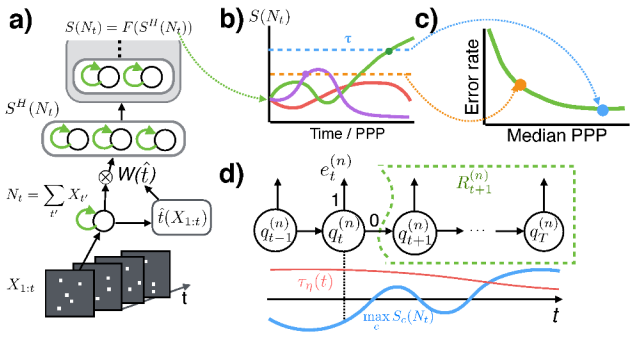

Figure 2. **WaldNet for lowlight visual recognition. (a)** Computing class posterior ratios. Time-invariant features $S^H(\boldsymbol{N}_t)$ from raw photon counts $\mathbf{X}_{1:t}$ are adapted based on exposure time t assuming constant illuminance (**Eq. 4**). If the illuminance is irregular and/or unknown, an equivalent exposure time $\hat{t}$ is estimated directly from the photon counts (**Sec. 3.5**). The first layer features feed into the remainder of the ConvNet F to compute class posterior $S_c(\boldsymbol{N}_t) = \frac{P(C=c|\mathbf{X}_t)}{P(C\neq c|\mathbf{X}_t)}$. $S_c(\boldsymbol{N}_t)$ may be computed efficiently using a spiking recurrent neural network implementation (**Sec. 3.6**) that leverages sparsity in changes of the network's internal states. **(b)** Deciding when to stop collecting photons. The class posterior ratios race to a common threshold to determine the category to report. WaldNet stops photon collection as soon as one class crosses the threshold τ (**Eq. 3**). A high threshold (blue) yields a later but more accurate solution whereas a lower threshold (orange) is faster but risks misclassification. **(c)** A speed versus accuracy tradeoff curve (illustration only) produced by repeating (a-b) for multiple images and sweeping the threshold τ.**(d)** Time-varying threshold $\tau_\eta(t)$ is trained to optimize Bayes risk with time cost η (**Eq. 5**). 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 $t+1$ (see **Eq. 7** for details).

of SPRT is at most $1-Sigm(\tau)$, where $Sigm$ is the sigmoid function: $Sigm(x) \triangleq \frac{1}{1+\exp(x)}$.

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 [44]. In the more realistic case where the categories are rich in intra-class variations, SPRT is shown to be asymptotically optimal, i.e. it gives optimal error rates as the exposure time becomes large [26]. Empirical studies suggest that even for short exposure times SPRT is near-optimal [6].

In essence, SPRT decides when to respond dynamically, based on the stream of observations accumulated so far. Therefore, the response time **varies** for each example. This regime is called "**free-response**" (FR), in contrast to the "**interrogation**" (INT) regime, typical of photopic vision, where each trial collects a **fixed-length** observation [4]. 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) is provably optimal in the asymptotical tradeoff between response time and error rate, while such proofs do not exist for the INT regime. We will empirically evaluate both regimes in **Sec. 4.3.1**.

3.3. Computing class probabilities over time

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

3.3.1 Model-free approach: network ensembles

The simple idea is to build a separate 'specialist' classifier $S(\boldsymbol{N}_t)$ 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 an 'ensemble' of specialists for a list of representative light levels, and route input counts $\boldsymbol{N}_t$ to the closest specialist in terms of light level.

One potential drawback of this ensemble approach is that training and storing multiple specialists 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 a more parsimonious system.

3.3.2 Model-based approach: WaldNet

Unlike the ensemble approach, we show that one can exploit the structure of the input and build one system for images at all PPPs. The variation in the input $\boldsymbol{N}_t$ has two independent sources: 1) the stochasticity in the photon arrival times and 2) 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. 2b-c**) that combines ConvNets with SPRT. **WaldNet automatically adjusts the parameters of a ConvNet according to the exposure time** t. Thus a WaldNet may be viewed as an ensemble of infinitely many specialists that occupies the size of only one specialist.

The key ingredient to SPRT is the log posterior ratios $S(\boldsymbol{N}_t)$ over exposure time t. Standard techniques such as

ConvNets can not be applied directly as they operate on a static input $\boldsymbol{N}_T$, the cumulative photon counts up to an identical exposure time T (e.g. $T\Delta \approx 33ms$ in normal lighting conditions). However we propose a simple adjustment that transfers the uncertainty in the photon counts to uncertainty in the task-relevant features of a ConvNet.

Standard ConvNets contain 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[5], and (2) the remaining layers $S(\boldsymbol{N}_T) = F(S^H(\boldsymbol{N}_T))$ that map the features $S^H(\boldsymbol{N}_T)$ to the log class posterior probabilities $S(\boldsymbol{N}_T)$. $\mathbf{W} \in \mathbb{R}^{D\times n_H}$ and $\mathbf{b}^H \in \mathbb{R}^{n_H}$ are the weights and biases.

Given only partial observations $\boldsymbol{N}_t$, computing features of the first layer requires marginalizing out unobserved photon counts $\Delta\boldsymbol{N} \triangleq \sum_{t'=t+1}^{T} \mathbf{X}_{t'}$. The marginalization requires putting a prior on the photon emission rate per image pixel i, which we assume to be a Gamma distribution: $Gam(\mu_i t_0, t_0)$, where μ_i represents the prior mean rate for pixel i and t_0 (shared across pixels) represents the strength of the prior[6]. Then the first layer of hidden features may be approximated by:

$$S^H(\boldsymbol{N}_t) \approx \alpha(t)\boldsymbol{W}\boldsymbol{N}_t + \boldsymbol{\beta}(t) \tag{4}$$

where the scaling factor $\alpha(t) \triangleq \frac{T+t_0}{t+t_0}$ is a scalar and the biases $\boldsymbol{\beta}(t)$ is a length n_H vector. For the j-th hidden feature, $\beta_j(t) \triangleq \frac{t_0(T-t)}{t+t_0}\sum_i W_{ij}\mu_i + b_j$. Derivations are in **Sec. A.1**.

Essentially, the adaptation procedure in **Eq. 4** accounts for the stochasticity in photon arrival time by using time-dependent weights and biases, rendering an exposure-time invariant feature representation $S^H(\boldsymbol{N}_t)$. The computations downstream, F, may then treat $S^H(\boldsymbol{N}_t)$ as if it were obtained from the entire duration. Therefore the same computations suffice to model the intra- and inter-class variations: $S(\boldsymbol{N}_t) = F(S^H(\boldsymbol{N}_t))$.

The network is trained discriminately (**Sec. 3.4**) with the first layer replaced by **Eq. 4**. The 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**. See **Sec. A.6** for implementation details.

3.4. Learning

Our goal is to train WaldNet to optimally trade off the expected exposure time (or PPP) and error rate in the FR regime. Optimality is defined by the Bayes risk R [44]:

$$R \triangleq \eta\mathbb{E}[\text{PPP}] + \mathbb{E}[C \neq \hat{C}] \tag{5}$$

where $\mathbb{E}[\text{PPP}]$ is the expected (over examples) photon count required for classification, $\mathbb{E}[C \neq \hat{C}]$ is the error rate, and η describes the user's cost of photons per pixel (PPP) versus error rate. WaldNet asymptotically optimizes the Bayes risk provided that it can faithfully capture the class log posterior ratio $S_c(\boldsymbol{N}_t)$, and selects the correct threshold τ (**Eq. 3**) based on the tradeoff parameter η. Sweeping η traverses the optimal time versus error tradeoff (**Fig. 2c**).

Since picking the optimal threshold according to η is independent from training a ConvNet to approximatel the log posterior ratio $S_c(\boldsymbol{N}_t)$, the same ConvNet is shared across multiple η's. This suggests the following two-step learning algorithm.

Step one: posterior learning

Given a dataset $\{\boldsymbol{N}_t^{(n)}, C^{(n)}\}_{n,t}$ where n indexes training examples (i.e. photon streams) and t denotes exposure times, we train the adapted ConvNet to minimize:

$$-\sum_{n,t} \log P(\hat{C} = C^{(n)} | \boldsymbol{N}_t^{(n)}) \tag{6}$$

When a lowlight dataset is not available we simulate the dataset from normal images according to the noise model in **Eq. 1**, where the exposure times are sampled uniformly on a logarithmic scale (see **Sec. 4**).

Step two: threshold tuning

If the ConvNet in step one captures the log posterior ratio $S_c(\boldsymbol{N}_t)$, we can simply optimize a scalar threshold τ_η for each tradeoff parameter η. In practice, we may opt for a time-varying threshold $\tau_\eta(t)$ for calibration purposes [7]. $\tau_\eta(t)$ affects our Bayes risk objective in the following way (**Fig. 2d**). Consider a high-quality (i.e. $T \to \infty$) image $\boldsymbol{N}_T^{(n)}$, let $\{\boldsymbol{N}_t^{(n)}\}_{t=1}^T$ be a sequence of lowlight images increasing in PPP generated from $\boldsymbol{N}_T^{(n)}$. Denote $q_t^{(n)} \triangleq \mathbb{I}[\max_c S_c(\boldsymbol{N}_t^{(n)}) > \tau_\eta(t)]$ the event that the log posterior ratio 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 incurred from time t onwards. $R_t^{(n)}$ may be computed recursively (derived in **Sec. A.3**):

$$R_t^{(n)} = \eta\Delta + q_t^{(n)}e_t^{(n)} + (1 - q_t^{(n)})R_{t+1}^{(n)} \tag{7}$$

where the first term is the cost of collecting photons at time t, 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.

[5] Without loss of generality and for notational simplicity, we assume that the first layer is fully-connected as oppose to convolutional. **Sec. A.1** discusses how to extend the results here to convolutional layers. We also define the first layer feature as the activity prior to non-linearity.

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

[7] This is because the learning in step-one can recover the $S_c(\boldsymbol{N}_t)$ up to a scaling and offset for each exposure time. The time-varying thresholds help to normalize the scales and offsets across time.

The Bayes risk, our minimization objective, is obtained from averaging multiple photon count sequences, i.e. $R = \mathbb{E}_{(n)}[R_0^{(n)}]$. To make R differentiable we approximate the non-differentiable component $q_t^{(n)}$ with a Sigmoid function $Sigm\left(\frac{1}{\sigma}(\max_c S_c(\boldsymbol{N}_t^{(n)}) - \tau_\eta(t))\right)$, and anneal the temperature σ over the course of training [30] (see **Sec. A.6**).

3.5. Automatic light-level estimation

Both scotopic algorithms (ensemble and WaldNet) assume knowledge of the light-level PPP in order to choose the right model parameters. This knowledge is easy to acquire when the illuminance is constant over time because PPP is linear in the exposure time t (**Eq. 2**), which may be measured by an internal clock.

However, in situations where the illuminance is dynamic and unknown, the linear relationship between PPP and exposure time is lost. In this case we propose to estimate PPP directly from the photon stream itself, as follows. Given a cumulative photon count image $\boldsymbol{N}$ (t, the time it takes to accumulate the photons, is no longer relevant as the illuminance is unknown), we examine local neighborhoods that receive high photon counts, and pool the photon counts as a proxy for PPP. In detail, we (1) convolve $\boldsymbol{N}$ using an $s \times s$ box filter, (2) compute the median of the top k responses, and (3) fit a second order polynomial to regress the median response towards the true PPP. Here s and k are parameters, which are learned from a training set consisting of ($\boldsymbol{N}$,PPP) pairs. Despite its simplicity, this estimation procedure works well in practice, as we will see in **Sec. 4**.

3.6. Spiking implementation

One major challenge of scotopic systems is to compute log posterior ratio computations as quickly as photons stream in. Photon-counting sensors [36, 12] sample the world at $1k - 10kHz$, overshadowing the fastest reported throughput of ConvNets [35][8]. Fortunately, since the photon arrival events within any time bin is sparse, changes to the input and the internal states of a scotopic system are small. We thus propose an efficient implementation that models only the changes above a certain magnitude. Our implementation relies on spiking recurrent hidden units:

1) the first hidden layer of WaldNet $S^H(\boldsymbol{N}_t)$ may be computed using the recurrent dynamics:

$$S^H(\boldsymbol{N}_t) = r(t)S^H(\boldsymbol{N}_{t-1}) + \alpha(t)\boldsymbol{W}\mathbf{X}_t + \boldsymbol{l}(t) \qquad (8)$$

where $r(t) \triangleq \frac{\alpha(t)}{\alpha(t-1)}$ is a damping factor, $\boldsymbol{l}(t) \triangleq \boldsymbol{\beta}(t) - r(t)\boldsymbol{\beta}(t-1)$ is a leakage term (derivations in **Sec. A.4**). The photon counts $\mathbf{X}_t$ in $[(t-1)\Delta, t\Delta)$ is sparse, thus the computation $\boldsymbol{W}\mathbf{X}_t$ is more efficient than computing $S^H(\boldsymbol{N}_t)$ from scratch.

[8]The throughput is $2kHz$ for 32×32 color images and $800Hz$ for 100×100 color images

2) We only propagate a change either in the positive or the negative direction when the its magnitude exceeds a predefined discretization threshold τ_{dis}.

3) Internal layers are represented using recurrent dynamics and discretized the same way.

The threshold affects not only the number of communication spikes, but also the quality of the discretization, and in turn the classification accuracy. For spike-based hardwares the number of spikes is an indirect measure of the energy consumption (Fig. 4(B) of [29]). For non-spiking hardwares, the number of spikes also translate to the number of floating point multiplications required for the layers above. Therefore, the τ_{dis} controls the tradeoff between accuracy and power / computational cost. We will empirically evaluate this tradeoff in **Sec. 4**.

4. Experiments

4.1. Baseline Models

We compare WaldNet against the following baselines, under both the INT regime and the FR regime. **For the first three baselines we assume that an internal clock measures the exposure time** t, and the illuminance λ^ϕ is known and constant over time.

1) **Ensemble**. We construct an ensemble of 4 specialists with PPPs from $\{.22, 2.2, 22, 220\}$ respectively. The performance of the specialists at their respective PPPs gives a upper bound on the optimal performance by ConvNets of the same architecture.

2) **Photopic classifier**. An intuitive idea is to apply a network trained in normal lighting conditions to properly rescaled lowlight images. 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.

3) **Rate classifier**. A ConvNet on the time-normalized image (rate) without weight adaptation. The first hidden layer is computed as $S_j^H(\boldsymbol{N}_t) \approx \boldsymbol{W}\boldsymbol{N}_t/t + \boldsymbol{b}^H$. Note the similarity with the WaldNet approximation used in **Eq. 4**.

4) **WaldNet with estimated light-levels (EstP)**. A WaldNet that is trained on constant illuminance λ^ϕ, but tested in environments with unknown and dynamic illuminance. In this case the linear relationship between exposure time t and PPP (**Sec. 2**) is lost. Instead, the light-level is first estimated according to **Sec. 3.5** directly from the photon count image $\boldsymbol{N}$. The estimate $\hat{PP}$ is then converted to an 'equivalent' exposure time $\hat{t}$ using $\hat{t} = \frac{\hat{PP}}{\lambda^\phi \Delta}$ (by inverting **Eq. 2**), which is used to adapt the first hidden layer of WaldNet in **Sec. 4**, i.e. $S^H(\boldsymbol{N}) \approx \alpha(\hat{t})\boldsymbol{W}\boldsymbol{N} + \boldsymbol{\beta}(\hat{t})$.

4.2. Datasets and Training

We consider two standard datasets: MNIST [22] and CIFAR10 [20]. We simulate lowlight training image sequences using **Eq. 1** and testing photon streams using the noise model of photon-counting sensors [12]. We set dark current

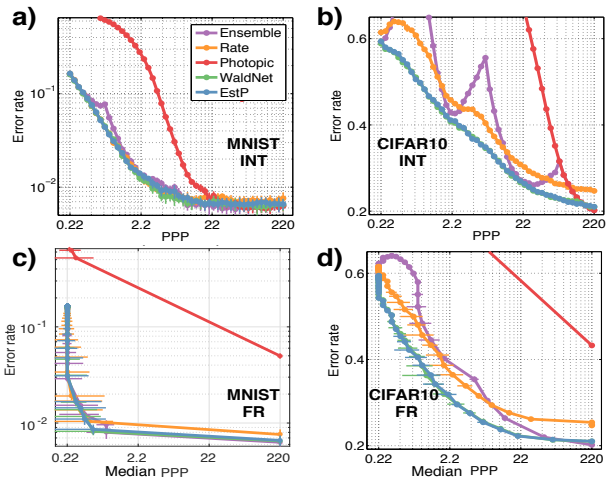

Figure 3. **Performance comparison**. (**a,b**) Error rate vs. the interrogation PPP for (**a**) MNIST and (**b**) CIFAR10. Each dot is computed from classifying $10k$ test examples with a fixed PPP. (**c,d**) Error rate vs. *median* PPP for (**c**) MNIST and (**d**) CIFAR10 (In FR regime, every photon stream requires a different PPP to be classified, hence the median is used.). 1 bootstrap *ste* is shown for both the median PPP and error rate, the latter is too small to be visible.

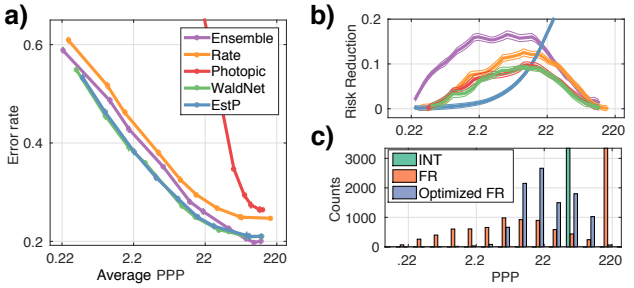

Figure 4. **The effect of threshold learning (Sec. 3.4)**. (**a**) Error rate vs. *average* PPP for CIFAR10 using a network with optimized time-varying threshold $\tau_\eta(t)$. 1 bootstrapped *ste* is shown but not visible. (**b**) Each curve shows the Bayes risk reducation after optimization (**Sec. 3.4**, step 2) per *average* PPP. (**c**) Response time (PPP) histograms under INT, FR before, and FR after optimization of a WaldNet that achieves 22% error on CIFAR10.

$\epsilon_{dc} = 3\%$ and ignore other noise sources for model comparison in **Sec. 4.3**, and separately evaluate the effect of all noise sources in **Sec. 4.4**. We use the default LeNet [22] for MNIST and the CIFAR10-quick architecture from the MatConvNet package [42], both with batch normalization [37] and without data augmentation. Details of the models and training are described in **Sec. A.5** and **A.6**, and the code (based on MatCovNet [42]) is online[1].

4.3. Results

The speed versus accuracy tradeoff curves in the INT regime are shown in **Fig. 3a,b**. Median PPP versus accuracy tradeoffs for all models in the FR regime are shown in **Fig. 3c,d**. All models use constant thresholds for producing the tradeoff curves. In **Fig. 4a** are average PPP ver-

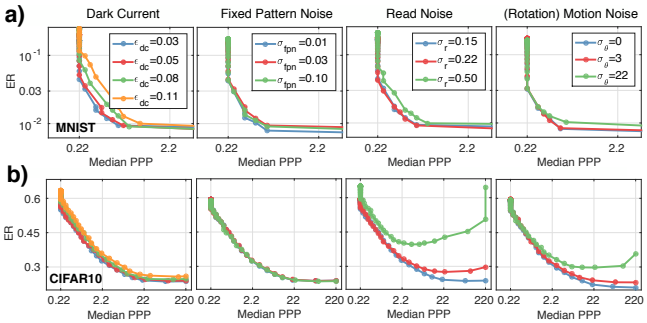

Figure 5. **The effect of sensor noise on WaldNet**. The rows correspond to datasets MNIST and CIFAR10, and the columns correspond to parameters of noise sources, which are the dark current ϵ_{dc}, the standard deviation of multiplicative fixed pattern noise σ_{fpn}, the std of additive read noise σ_r, and the std of the rotational jitter σ_θ in degrees. Only one noise is varied in each panel while the rest are fixed at their respective baseline: $\epsilon_{dc} = 3\%$,$\sigma_r = 0.15, \sigma_{fpn} = 3\%$ and $\sigma_\theta = 0$.

sus accuracy curves when the models use optimized dynamic thresholds described in **Sec. 3.4**, step-two.

4.3.1 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.

WaldNet is sufficient. The ensemble was formed using specialists at logarithmically-spaced exposure times, thus its curve is discontinuous in the interrogation regime (esp. **Fig. 3b**). The peaks delineate transitions between specialists. Even though WaldNet uses $1/4$ the parameters of the ensemble, it stays close to the performance upper bound (estimated from ensemble performance). Under the FR regime, WaldNet is indistinguishable from the ensemble in MNIST and superior to the ensemble in lowlight conditions ($\leq$ 22 PPP) of CIFAR10.

Training with scotopic images is necessary. The photopic classifier retrofitted to lowlight applications performs well at high light conditions ($\geq$ 220 PPP) but works poorly overall in both datasets. Investigation reveals that the classifier fails to assess probability of low light images and often stops evidence collection prematurely.

Weight adaptation is necessary. The rate classifier slightly underperforms WaldNet in both datasets. Since the two system have the same degrees of freedom and differ only in how the first layer feature is computed, the comparison highlights the advantage of time-adaptation (**Eq. 4**).

FR is better than INT. Cross referencing **Fig. 3a,b** and **Fig. 3c,d** reveals that FR with constant thresholds often brings 3x reduction in median photon counts.

4.3.2 Effect of threshold learning

The comparison above under the FR regime uses constant thresholds on the learned log posterior ratios (**Fig. 3c,d**). Us-

ing learned dynamic thresholds (step two of **Sec. 3.4**) we see consistent improvement on the *average* PPP required for given error rate across all models (**Fig. 4b**), with more benefit for the photopic classifier. **Fig. 4c** 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 reduces the variance of the PPP but make the median longer. This is ok because the Bayes risk objective (**Eq. 5**) 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**.

4.4. Sensitivity to sensor noise

How robust is the speedup observed in **Sec. 4.3** affected by sensor noise? For MNIST and CIFAR10, we take WaldNet and vary independently the dark current, the read noise, the fixed pattern noise and a rotational jitter noise where a random rotation parameterized by σ_θ is applied per unit time (details in **Sec. A.7**).

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 ($\sigma_r = 22\%$ [12]) guarantees no performance loss for MNIST and minor loss for CIFAR10, suggesting the need for improvement in both the algorithm and the photon-counting sensors. The fact that $\sigma_r = 50\%$ hurts performance also suggests that single-photon resolution is vital for scotopic vision. Lastly, while WaldNet provides certain tolerance to rotational jitter, drastic movement (22° at 220 PPP) could cause significant drop in performance, suggesting that future scotopic recognition systems and photon-counting sensors should not ignore camera / object motion.

4.5. Efficiency of spiking implementation

Finally, we inspect the power efficiency of the spiking network implementation (**Eq. 8**) on the MNIST dataset. We assume that a photon-counting sensor observes a photon stream totalling 22 PPP, and reports 100 binarized "images" of photon arrival events (i.e. at .22 PPP / frame). Our reference implementation ("Continuous") runs a ConvNet from end-to-end every time a binary image arrives. The spiking implementations employ a discretization threshold τ_{dis} (**Sec. 3.6**) that is common across all layers. As a proxy for power efficiency we use the number of multiplications and additions (MultAdds) [29], normalized by the MultAdds of running the baseline throughout the whole duration.

The power vs accuracy tradeoff results in **Fig. 6a,b** suggest that the discretization threshold is optimal near $\tau_{dis} = 0.2$, where the spiking implementation is $2-3\times$ more efficient than the continuous implementation at equal error rate in FR mode (**Fig. 6a**). **Fig. 6b** suggests that the spiking implementation becomes increasingly more efficient over time as the network's signals become more stable, and the spiking implementation with $\tau_{dis} = 0.2$ is $5\times$ more efficient than the baseline in INT mode.

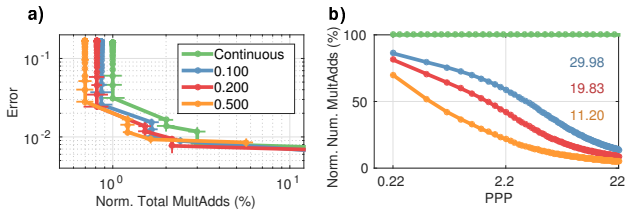

Figure 6. **"Energy" and accuracy tradeoff of the spiking recurrent neural network implementation on MNIST**. **a)** Error rates of spiking networks with different discretization thresholds τ_{dis} (**Eq. 8**) against the MultAdds (normalized) running in FR mode. **b)** MultAdds per frame (normalized) as a function of time / PPP. Numbers inset represent the cumulative percentage of MultAdds of the network running in INT mode till PPP$=22$ (normalized).

5. Discussion and Conclusions

We study the important yet relatively unexplored problem of scotopic visual recognition, where the available light is low or expensive to acquire, and image capture is more lengthy / costly than computation. 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.

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 amount of photons required for decision at desired accuracy). WaldNet further allows for a flexible tradeoff between energy / computational efficiency with accuracy when implemented as a recurrent spiking network. When trained assuming a constant illuminance, WaldNet may be applied in environments with varying and unknown illuminance levels. Finally, despite relying only on few photons for decisions, WaldNet is minimally affected by camera noises, making it ideal for integration with the evolving lowlight sensors.

References

[1] Scotopic vision on github. https://github.com/bochencaltech/scotopic. 7

[2] L. Anzagira and E. R. Fossum. Color filter array patterns for small-pixel image sensors with substantial cross talk. *JOSA A*, 32(1):28–34, 2015. 3

[3] H. Barlow. A method of determining the overall quantum efficiency of visual discriminations. *The Journal of physiology*, 160(1):155–168, 1962. 2

[4] 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*, 113(4):700, 2006. 4

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

[6] B. Chen and P. Perona. Towards an optimal decision strategy of visual search. *arXiv preprint arXiv:1411.1190*, 2014. 4

[7] B. Chen and P. Perona. Scotopic visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 8–11, 2015. 2

[8] B. Chen, P. Perona, and L. D. Bourdev. Hierarchical cascade of classifiers for efficient poselet evaluation. In *BMVC*, 2014. 2

[9] T. Delbruck. Silicon retina with correlation-based, velocity-tuned pixels. *Neural Networks, IEEE Transactions on*, 4(3):529–541, 1993. 2

[10] T. Delbrück and C. Mead. Analog vlsi phototransduction. *Signal*, 10(3):10, 1994. 2

[11] E. D. Dickmanns. *Dynamic vision for perception and control of motion*. Springer Science & Business Media, 2007. 2

[12] E. Fossum. The quanta image sensor (qis): concepts and challenges. In *Imaging Systems and Applications*, page JTuE1. Optical Society of America, 2011. 2, 3, 6, 8

[13] E. R. Fossum. Modeling the performance of single-bit and multi-bit quanta image sensors. *Electron Devices Society, IEEE Journal of the*, 1(9):166–174, 2013. 3

[14] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980. 2

[15] 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. 2

[16] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):855–868, 2009. 2

[17] G. E. Healey and R. Kondepudy. Radiometric ccd camera calibration and noise estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(3):267–276, 1994. 3

[18] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 2

[20] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images, 2009. 6

[21] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1106–1114, 2012. 2

[22] 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. 2, 6, 7

[23] X.-D. Li, J. K. Ho, and T. W. Chow. Approximation of dynamical time-variant systems by continuous-time recurrent neural networks. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 52(10):656–660, 2005. 2

[24] 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. 3

[25] S.-C. Liu, T. Delbruck, G. Indiveri, A. Whatley, and R. Douglas. *Event-based Neuromorphic Systems*. John Wiley & Sons, 2014. 2

[26] G. Lorden. Nearly-optimal sequential tests for finitely many parameter values. *The Annals of Statistics*, pages 1–21, 1977. 2, 4

[27] D. C. Martin, D. Chang, M. Matuszewski, P. Morrissey, S. Rahman, A. Moore, and C. C. Steidel. Intergalactic medium emission observations with the cosmic web imager. i. the circum-qso medium of qso 1549+ 19, and evidence for a filamentary gas inflow. *The Astrophysical Journal*, 786(2):106, 2014. 2

[28] J. Matas and O. Chum. Randomized ransac with sequential probability ratio test. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1727–1732. IEEE, 2005. 2

[29] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197):668–673, 2014. 6, 8

[30] 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*, pages 43–56. Springer, 2015. 6

[31] P. Moreels, M. Maire, and P. Perona. Recognition by probabilistic hypothesis construction. In *Computer Vision-ECCV 2004*, pages 55–68. Springer, 2004. 2

[32] 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, 3

[33] M. Naghshvar, T. Javidi, et al. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738, 2013. 2

[34] P. O'Connor, D. Neil, S.-C. Liu, T. Delbruck, and M. Pfeiffer. Real-time classification and sensor fusion with a spiking deep belief network. *Frontiers in neuroscience*, 7, 2013. 2

[35] K. Ovtcharov, O. Ruwase, J.-Y. Kim, J. Fowers, K. Strauss, and E. S. Chung. Accelerating deep convolutional neural networks using specialized hardware. *Microsoft Research Whitepaper*, 2, 2015. 6

[36] 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*, pages 1093–1096. IEEE, 2009. 2, 3, 6

[37] C. S. Sergey Ioffe. Batch normalization: Accelerating deep network training by reducing internal covariate shift. volume 32, pages 448–456, 2015. 7

[38] D. J. Stephens and V. J. Allan. Light microscopy techniques for live cell imaging. *Science*, 300(5616):82–86, 2003. 2

[39] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 2

[40] A. G. Tartakovsky. Asymptotic optimality of certain multihypothesis sequential tests: Non-iid case. *Statistical Inference for Stochastic Processes*, 1(3):265–295, 1998. 2

[41] S. Thorpe, D. Fize, C. Marlot, et al. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996. 2

[42] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. 2015. 7

[43] 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*, volume 1, pages I–511. IEEE, 2001. 2

[44] A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945. 2, 4, 5

[45] M. N. Wernick and G. M. Morris. Image classification at low light levels. *JOSA A*, 3(12):2179–2187, 1986. 2

[46] F. Zappa, S. Tisa, A. Tosi, and S. Cova. Principles and features of single-photon avalanche diode arrays. *Sensors and Actuators A: Physical*, 140(1):103–112, 2007. 2, 3

[47] M. Zhu, N. Atanasov, G. J. Pappas, and K. Daniilidis. Active deformable part models inference. In *European Conference on Computer Vision*, pages 281–296. Springer, 2014. 2