Learning Attention Models for Resource-Constrained, Self-Adaptive Visual Sensing Applications

Hafiz Areeb Asad
hafiz.a.asad@ntnu.no
Norwegian University of Science and Technology, NTNU
Trondheim, Norway

Frank Alexander Kraemer
kraemer@ntnu.no
Norwegian University of Science and Technology, NTNU
Trondheim, Norway

Kerstin Bach
kerstin.bach@ntnu.no
Norwegian University of Science and Technology, NTNU
Trondheim, Norway

Christian Renner
christian.renner@tuhh.de
Hamburg University of Technology, TUHH
Hamburg, Germany

Tiago Santos Veiga
tiago.veiga@ntnu.no
Norwegian University of Science and Technology, NTNU
Trondheim, Norway

ABSTRACT

Resource constraints are one of the main design challenges for wireless sensor network applications and visual sensing networks that employ cameras in particular. The objective in this paper is to enable the sensors to be context-aware by utilizing application-level information, to prioritize parts of an image, and only transmit those parts that contribute most to the utility of the application. We, therefore, study online-learning of visual attention models for the use case of person detection and counting. We analyze how the resulting models can prioritize relevant elements of a partial image, so that object detection remains accurate compared to a random selection strategy when resources for transmission get scarce. Results show that such attention models can be learned also under constraints and converge towards the true models. For the application performance, we observed an average reduction of errors (the number of undetected persons) of 55% compared to policies without a corresponding attention model.

KEYWORDS

adaptive sensing, crowd detection, internet of things, image processing, machine learning, online learning

ACM Reference Format:
Hafiz Areeb Asad, Frank Alexander Kraemer, Kerstin Bach, Christian Renner, and Tiago Santos Veiga. 2022. Learning Attention Models for Resource-Constrained, Self-Adaptive Visual Sensing Applications. In *International Conference on Research in Adaptive and Convergent Systems (RACS '22), October 3–6, 2022, Virtual Event, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3538641.3561505

1 INTRODUCTION

Visual sensor networks that use cameras as main detectors are interesting since they can be installed and utilized quite flexibly, and used for a wide range of use cases, like surveillance [12], environmental monitoring [5], or agriculture [6]. The problem is that they can be resource-intensive in terms of transmission bandwidth, which can be a challenge both for the energy consumption of the individual sensor as well as the load of the wireless channels in the network [5, 15]. Efficient operation through low-power computation, optimized protocols and energy harvesting [14] are technologies to facilitate these constraints at a lower level. Transmissions may also be reduced through compression [12] or change detection [6] directly at the image level.

In addition to such lower-level optimizations, the application level adds possibilities to align the operation of the device with its actual utility. Given the specific application goals, not all observations are equally useful, and devices can benefit from the ability to prioritize data, also referred to as value-of-information [1]. The precondition for such frugal, self-aware operation is knowledge within the device about its own operation, the environment, and the goals of the application. With such knowledge, devices can use their resources more strategically on data of high value and are hence better prepared to maintain their utility also in constrained situations. Such approaches are also referred to as self-adaptive [2, 11], context-aware or cognitive IoT [4].

In this paper, we follow such a self-adaptive and cognitive approach and present a visual sensing network with the specific use case of detecting persons in a skiing area to estimate its busyness and utilization. We want to find out to which degree we can learn and utilize visual attention models. These models capture which parts of an image are relevant for people counting, so that devices can decide which parts of an image to transmit once transmission needs to be restricted. The key challenges with such an approach are that (1) sensors are plentiful and deployed in heterogeneous settings which require individual adaptation, and learning must therefore happen autonomously; (2) learning needs to start without prior knowledge, as data only becomes available after deployment; (3) learning happens also during resource-constrained operations, that means, it must be possible to also learn with reduced data transmissions.

We, therefore, describe and study the effect of a visual attention model that partitions a camera image into discrete tiles, and that captures which tiles are relevant for the detection of persons by

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
RACS '22, October 3–6, 2022, Virtual Event, Japan
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9398-0/22/10.
https://doi.org/10.1145/3538641.3561505

analyzing the bounding boxes from an image recognition algorithm. Results show that the attention models acquired through reduced transmission levels converge with the true values, and that utilizing attention models can improve the accuracy of the applications considerably, when compared to policies with similar reductions that do not use the knowledge from the attention models.

The remainder of this paper is organized as follows. Section 2 provides a brief account of state-of-the-art solutions that are related to our work. In Section 3 we present the basic use case, while in Section 4 we develop the concept of visual attention and how it is learned and acted upon. Section 5 evaluates the results of our solution.

The complete code of our experiments is publicly available.[1]

2 RELATED WORK

The approach to use constrained resources strategically on information items that are valuable is also referred to as value-of-information and finds its application in a wide range of use cases in IoT and wireless sensor networks, see e.g., [1]. To avoid inefficient operation due to static parameter settings in changing environments, numerous approaches have been proposed in the literature to make IoT systems context-aware and self-adaptive [2, 11]. These approaches mostly employ the MAPE (Monitor-Analyze-Plan-Execute) mechanism for self-adaptation in different tiers of the system like sensor, communication layer or cloud.

For real-time visual surveillance applications, it is desired that the transmission cost of sending continuous images for monitoring a static environment should be minimal. For example, Ji *et al.* presented a scheme based on dissimilarity measure to transmit only parts of the modified image for the application of wide-area agriculture farms [6]. The dissimilarity measure compares the current image with the previous image and identifies changed patches to send. In this way, the multimedia size is reduced and the required bandwidth for transmission via IoT communication protocols is respected. However, this requires an application-specific dissimilarity threshold.

Similarly, to relax channel occupancy in Narrow Band (NB)-IoT, Khan *et al.* [8] present an approach to run object detection on a gateway and send only detected parts. This follows a similar principle of utilizing information about object locations as we do, but does not learn an attention model as we do for this self-adaptive approach. In addition, several lightweight image transmission and compression techniques via LoRa technology have been proposed to address bandwidth and power limitations [15].

To avoid the transmission of images entirely, they could also be processed directly on the sensor device. Some object detection and classification has recently become possible on low-power embedded systems [10]. However, this approach currently only handles objects of similar size, and are not as flexible as we need for our use case, where persons can appear in various parts of the image, stand in groups or have different sizes. When these approaches become more capable, they could potentially benefit from learning an attention model as we explore here.

[1] https://github.com/areebasad/Learning-Attention-Models-for-Resource-Constrained-Self-Adaptive-Visual-Sensing-Applications

Figure 1: Two persons detected by the YOLO image recognition algorithm with corresponding confidence

Complementing these early works, in this paper, our goal is to present a self-adaptive approach for resource-constrained IoT systems that can learn from the environment and can adapt according to the changes and utilize on-board resources efficiently, further to maintain the device utility.

3 BASIC PERSON DETECTION

The application in our case study estimates the busyness of a skiing area by detecting and counting persons on images captured by cameras. Such data can be used to estimate the demand for public transport, parking, or the opening hours of restaurants, among others. The basis for our case study is data from five cameras deployed at different locations in a cross-country skiing area in Trondheim, Norway. Images are taken at a fixed interval of 10 minutes.

The person detection is done using a deep neural network for general object detection. It follows the *You only look once* approach (YOLO) [7], which can detect several objects in a single inference step. The benefit of such a general detection model is that it can detect persons of various sizes, postures and locations within an image, so that the camera devices need no particular strategic placement, which, in turn, simplifies deployment. The output of the object detection model is a set of bounding boxes, associated with the detected object (like *person*) and a confidence level. Figure 1 shows the result of the object detection, depicting recognition of two persons with their confidence scores. Experience shows that confidence levels above 0.5 provide good results for person counting in this context. For our system, we use a YOLO model (version 5s) provided through the AI4EU platform [13], which was trained from a general data set [9]. Hence, such a solution can be deployed and work without expensive training of the models. YOLO can work on different image sizes, we applied it to 384×640 pixels.

The diagrams in Fig. 2 show the results of the person counting in our data set. The upper graph shows the counted persons over a period of two weeks, the lower graph within a single day.

4 VISUAL ATTENTION MODELS

As outlined in the introduction, wireless devices offer more flexibility during deployment and lower installation costs, so that one could provide more fine-grained data through more monitoring points in the skiing area. However, they come with the challenge of resource constraints especially related to transmission [6]. Though

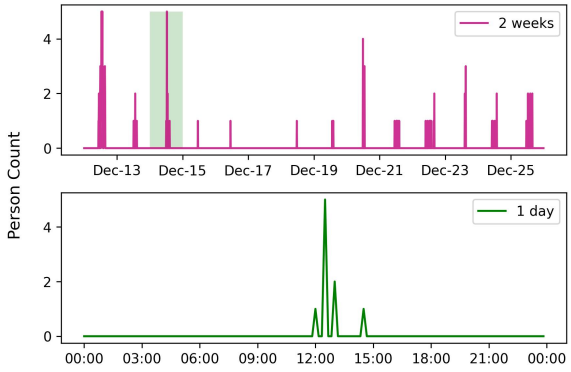

Figure 2: Person counts over two weeks (top) and within a day (bottom) extracted from the YOLO image detection

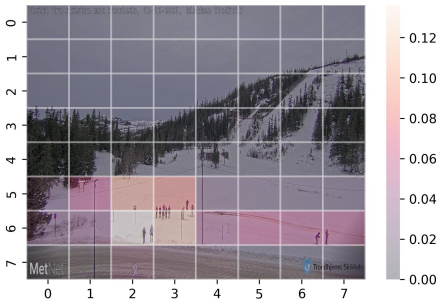

Figure 3: Learned visual attention model shown as a heatmap over the background scene

there are advances in embedded machine learning (see Sect. 2), capabilities are still limited when comparing with the state-of-the-art YOLO models. We therefore execute the actual person detection on a server, and the objective is to reduce the transmission of data. For that, we here want to focus on the selection of image parts at the application level. (Approaches for data compression or change detection can come in addition to that, but are out of scope here.)

For the detection and counting of persons, sending the entire image is in most cases not necessary. In Fig. 1, for example, it is unlikely to detect skiers in the sky or in areas covered by trees. The idea is hence to let the sensor devices learn which parts of the image are significant for person counting in the form of a visual attention model V (see, e.g., [3]), and use that knowledge to only send those parts of an image when resources are scarce. We therefore divide an image into N parts, referred to as *tiles*, illustrated in Fig. 3. The figure indicates through coloring the relative number of persons detected within each tile over time. As one can expect, tiles that cover the skiing tracks are more populated and hence more relevant for person counting than others. Figure 4 shows the relevance of the individual tiles for all of the cameras, based on the entire data set of 80 days. It is visible that distributions are specific to each camera, confirming that such models should be learned individually.

The system works as illustrated in Fig. 5. Images taken by the cameras are tiled, and the sensor devices decide based on a policy

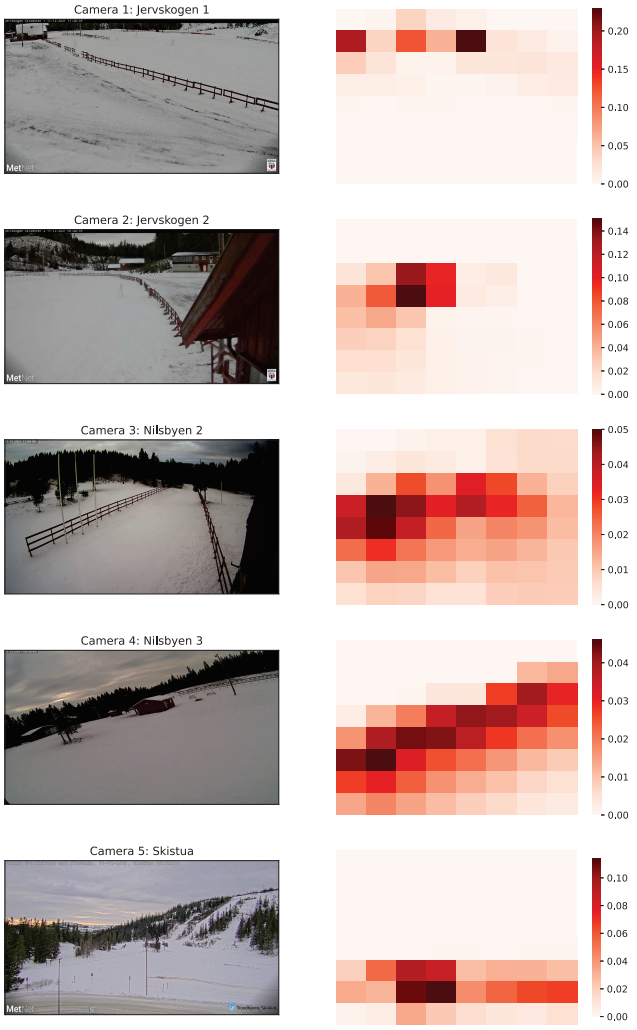

Figure 4: Camera views along with their visual attention model over the entire range of images captured

(explained in Sect. 4.2) how many and which of the tiles to send. Selected tiles are transmitted to the server, where they are combined into a complete image. For the missing parts of an image, the server simply selects a background image that was transmitted earlier. The server runs the object detection on the reconstructed image. The output is used for the person counting. In addition, the bounding boxes of the detection are used by the server to learn the visual attention models that are then transmitted to the sensor devices.

4.1 Learning the Visual Attention Models

The visual attention models in Fig. 4 were created from the entire data set as an illustration. In real deployments, the attention models must be learned online on the server, that means, while the system is running, from the incoming images, starting without prior information.

The attention model for a camera c is initialized with $V^c[t] \leftarrow 0 \; \forall \; t$, meaning all tiles t are considered equally relevant.

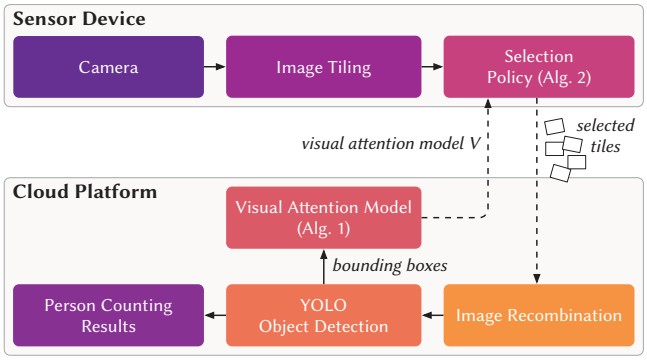

Figure 5: Architecture of the sensor device with camera and the cloud computing platform. Portions of images are sent based on the visual attention model

Algorithm 1: Updating the Visual Attention Model V

```
Data: bounding boxes with detections, current visual
      attention model V^c
p[t] ← 0 for all tiles t
forall bounding boxes b do
    forall tiles t do
        if b.covers(t) and b.confidence > 0.5 then
            p[t] ← p[t] + 1
        end
    end
end
for tiles t do
    V[t] ← α · p[t] + (1 − α) · V[t]
end
```

Algorithm 1 describes how to update the visual attention models over time. It takes a set of bounding boxes from the object detection as input, which can either come from a single image or a batch of images. The algorithm can hence run for each image received or for several images, for instance all received within one day. The algorithm uses a temporary variable p that counts the number of times any of the tiles t is covered by a bounding box. Once p registered all bounding boxes for all tiles, the attention model V is updated. Here we introduce a discount factor α, so that older observations weight less than more recent ones, similar to the effect of an exponentially weighted moving average (EWMA). This factor enables constant learning. Without α, values in V would monotonously increase, and newly detected persons would have less and less influence on the attention model. The model would not be able to adapt to changes in the environment anymore. Like with an EWMA, values for α can be estimated through the range of values to effectively take into account by $\alpha = 2/(range+1)$. For our experiment, we have chosen 30 days as appropriate time window, meaning that the attention map is mainly based on the number of persons within the last 30 days. With 144 images each day, this corresponds to $\alpha = 2/(144 \cdot 30 + 1) \approx 0.463 \times 10^{-3}$.

Algorithm 2: Policy P for tile selection

```
Data: visual attention model V^c, ϵ, transmission level l
with probability ϵ:
    sample n tiles with uniform weights
or with probability 1 − ϵ:
    sample n tiles weighted by V
```

4.2 Tile Selection Policy

The first question a policy has to answer is *how many tiles to transmit*, which has a direct influence on the resources used. A detailed cost analysis depends on the specific hardware and is out of scope for this paper; we proceed with the simplified assumption that fewer tiles mean fewer resources spent on transmission. To study the principles of the approach, we consider four constant transmission levels $l \in \{80\%, 60\%, 40\%, 20\%\}$, which for N=64 tiles correspond to the transmission of only $n = 51$, 38, 26, or 13 tiles, respectively.

We approach the problem of *which tiles to select* as a multi-armed bandit (see, e.g., [16]), where each action corresponds to the different subsets of tiles that can be transmitted. The problem is that the visual attention models are learned online, and only from the tiles we transmit. The attention model is hence changing over time, since we gradually learn more, and due to possible changes in the environment. If we only select the tiles that are marked as most interesting by the attention model, (we *exploit* its knowledge) we fail to *explore* other tiles and may not learn the true distribution of persons in the image. For that, Algorithm 2 lists an epsilon-based approach with a tradeoff between exploitation and exploration:

- If the policy selects to **exploit** (with probability $1 - \epsilon$), it selects tiles randomly but weighted by the probability distribution of the visual attention model.
- If the policy selects to **explore** (with probability ϵ), the policy selects tiles randomly but without considering the visual attention model, i.e., according to a uniform distribution.

We use an epsilon-decreasing strategy in which a device gradually reduces ϵ, that means, explores more at the start and increases exploitation over time. A policy P_l^c for camera c selects tiles according to transmission level l and the visual attention model V_l^c.

5 EVALUATION

To evaluate our approach, we simulated the operation of the system using the different policies P_l^c for each camera c and transmission level l, as if sensors would have been deployed, i.e., starting with zero knowledge and then simulating the operation on the entire set of images over 90 days. This is resource-intensive, since the YOLO object detection has to run on each instance of the reduced images anew, as each run may select different subsets of tiles. We selected $N = 64$ tiles and α corresponding to a time range of 30 days. The devices explore and exploit equally from the first day ($\epsilon = 0.5$) and gradually decrease exploration to 20% ($\epsilon = 0.2$) after 3 months. As ground truth, we define a policy P_*^c that transmits all tiles of an image.

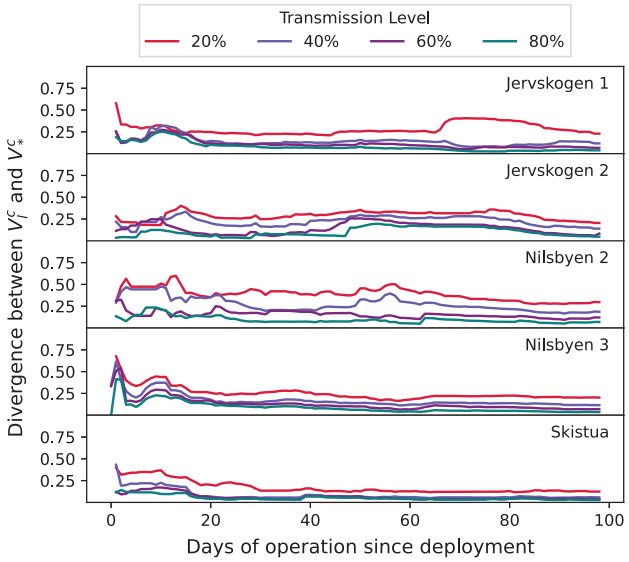

Figure 6: Divergence between attention models V_l^c and the ground truth V_*^c over time

5.1 Convergence of Visual Attention Models

The attention models V_l^c are computed from the images that were transmitted with the reduced levels l, and we want to investigate how the models diverge from the true models V_*^c that would represent the attention from images where all tiles are transmitted. Since the attention models are distributions over the tiles, we use the Jensen-Shannon distance metric which measures the similarity between two probability distributions P and Q:

$$\begin{aligned} \mathrm{JSD}(P \parallel Q) &= \frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M), \\ \textit{where } M &= \frac{1}{2}(P+Q) \textit{ and } D(X \parallel Y) = \sum X \log\left(\frac{X}{Y}\right) \end{aligned} \tag{1}$$

Values close to 1 indicate a high divergence, and values close to 0 that distributions are similar. Fig. 6 shows the divergence of the attention distribution learned with the reduced transmission levels when compared to the true distribution learned from the complete images. In general, we observe that the lower the transmission level, the longer it takes for the attention models to converge. We observe sporadic divergence, especially for lower transmission levels at locations *Jervskogen 1*, *Jervskogen 2* and *Nilsbyen 2*. A manual inspection revealed ski competitions in the corresponding images, which imply a significant change in the placement and number of persons. After such events, the attention models converge again. In *Nilsbyen 2*, the irregular raise is also amplified because of the wide coverage of persons at that area.

5.2 Distribution and Causes of Errors

We now analyze the errors that occur when we only transmit parts of the tiles with policies P_l^c utilizing the visual attention models V_l^c. For that, we compare for each image the number of persons detected using the policy with the ground truth. The error

	$l = 80\%$	$l = 60\%$	$l = 40\%$	$l = 20\%$
C1: Correct cases	96.4%	94.7%	93.1%	90.2%
C2: Cases w. undet. persons	2.2%	3.6%	5.3%	8.6%
C3: Cases w. false positives	1.3%	1.5%	1.5%	1.0%

Table 1: Frequency of the different result class for each transmitted image, all cameras combined

$e_{l,i}^c$ for an image i is then the difference between ground truth and the person count with the policy and reduced transmission, i.e., $e_{l,i}^c = \mathit{detected}(P_*^c, i) - \mathit{detected}(P_l^c, i)$. Fig. 7 shows the histogram of the errors for all transmission levels. We distinguish three cases, which are also detailed in Table 1.

C1 Images where the number of detected persons with the policy matches the ground truth.

This class marks cases where the tiling did not introduce any error, and Table 1 reveals that this is the vast majority of cases. They are represented in the histograms in Fig. 7 as the hatched bar at value 0. (It is off the charts due to the scale, which we selected so that the error classes are better visible.) In contrast, errors happen in the following two classes:

C2 Images where the reduced transmission leads to *undetected persons*. This results in a positive error, and is hence represented by the bars to the right of the hatched bar in Fig. 7. Not unexpected, this type of error is the most frequent one, since tiles not transmitted may contain persons that cannot be detected.

C3 Images where the reduced transmission leads to *more persons* detected than actually present, so-called false positives. These result in a negative error, and are shown to the left of the hatched bar. Fig. 8 illustrates such an error where the legs of a barrier were interpreted as a person when it was adjacent to a tile not transmitted.

There may be cases where both false positives and undetected persons happen within the same image, so that errors cancel each other out. There may also be errors in the ground truth, that means, the YOLO algorithm does not work correctly in the original image. This, however, is then an inherent challenge with the image recognition and not a problem of our tiling policy. From a random examination, however, we conclude that these cases do not happen often.

5.3 Value of the Attention Models

To determine to which degree selecting the tiles according to the visual attention model helps to prevent missed counts, we compare our policies P_l^c using the visual attention models with random policies R^l. These policies R^l use the same transmission levels l, but sample always randomly, without the knowledge of the visual attention model.

Table 2 shows the detection for all cameras, alongside the number of undetected persons for each transmission level. Camera *Jervskogen 1*, for instance, has a total of 4202 detected persons using the ground truth over the entire period of 90 days. With an 80%-transmission policy, the random policy missed 1434 persons, while the policy using the visual attention model only missed 385

			l = 80%		l = 60%		l = 40%		l = 20%	
Camera	True Count	Policy	Undetected	%	Undetected	%	Undetected	%	Undetected	%
Jervskogen 1	4202	Random R_l	1434	34.1%	2272	54.1%	3007	71.6%	3707	88.2%
		Learned P_l	385	9.2%	583	13.9%	1107	26.3%	2204	52.5%
		Difference	1049	**73.2%**	1689	**74.3%**	1900	**63.2%**	1503	**40.5%**
Jervskogen 2	3077	Random R_l	793	25.8%	1418	46.1%	1940	63.0%	2576	83.7%
		Learned P_l	263	8.5%	554	18.0%	780	25.3%	1351	43.9%
		Difference	530	**66.8%**	864	**60.9%**	1160	**59.8%**	1125	**47.6%**
Nilsbyen 2	3264	Random R_l	851	26.0%	1512	46.3%	2077	63.6%	2684	82.2%
		Learned P_l	331	10.1%	688	21.1%	1052	32.2%	1860	57%
		Difference	520	**61.1%**	824	**54.5%**	1025	**49.4%**	824	**30.7%**
Nilsbyen 3	6770	Random R_l	1604	23.7%	3024	44.7%	4167	61.6%	5535	81.8%
		Learned P_l	661	9.8%	1305	19.3%	2566	38%	4228	62.5%
		Difference	943	**58.8%**	1719	**56.8%**	1601	**38.4%**	1307	**23.6%**
Skistua	3753	Random R_l	984	26.2%	1651	44%	2461	65.6%	3230	86.1%
		Learned P_l	308	8.2%	631	16.8%	820	21.8%	1731	46.1%
		Difference	676	**68.7%**	1020	**61.8%**	1641	**66.7%**	1499	**46.4%**

Table 2: Residual errors in terms of undetected persons for each camera and transmission level l

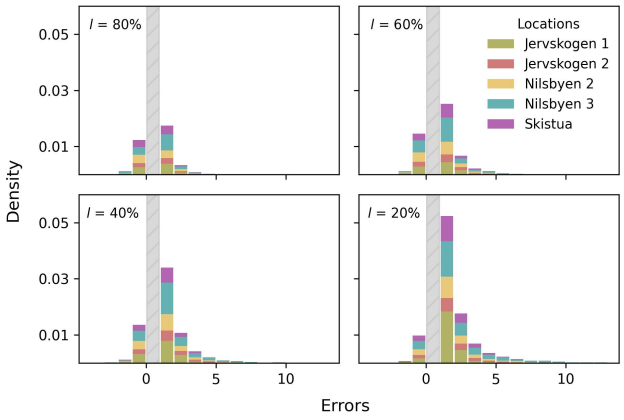

Figure 7: Distribution of errors for all transmission levels. The hatched bar depicts the correct estimation (zero error).

persons. This means the visual attention model prevented 73.2% of the errors that the random policy caused. Averaging over all cameras and transmission levels, the reduction in errors is 55%.

Fig. 9 visualizes the number of missed persons for each camera and each transmission level, and compares the policies using the visual attention models with the random policies. The vertical axis shows the percentage of persons missed relative to the total number. The plot shows clearly that for all cameras and transmission levels, the number of missed detections is greatly reduced for the policies utilizing the visual attention models.

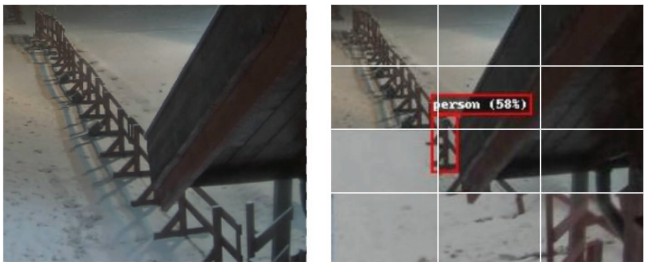

Figure 8: The full image (left) does not trigger a person detection, but tiling and combination with the background cause the fence to be interpreted as a person.

6 CONCLUSION AND OUTLOOK

We studied how visual attention models can be learned and exploited for the use case of person detection and counting in a skiing area. The attention models identify the most rewarding parts of an image and gain this knowledge over time, starting without prior knowledge. We have shown that the models can also be learned under resource constraints, and that their utilization can considerably reduce the number of undetected persons when transmissions need to be reduced.

In this study, we only considered policies with fixed transmission levels and an epsilon-decreasing strategy for exploration to establish the principle of the visual attention models. More advanced policies and attention models could take many more signals into account. Visual attention models could depend on time and other contextual information, and energy planners, which were out of

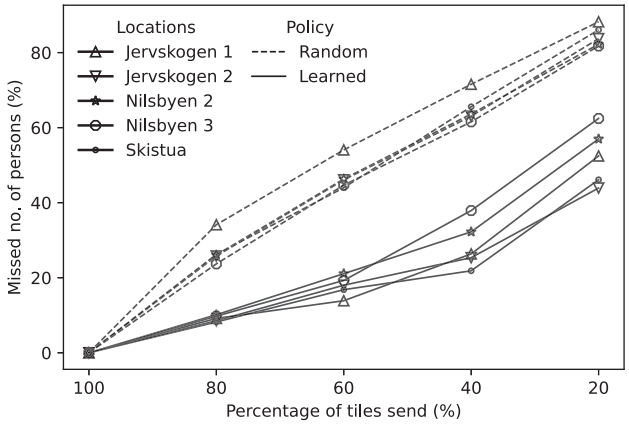

Figure 9: Percentages of undetected persons for each camera, using the visual attention models and for random policies

the scope in this paper, could determine the optimal trade-off between utilization and exploration for future exploitation taking the harvested energy and how it varies over time into account.

This use case fits into the larger picture of self-adaptive, cognitive IoT applications, in which low-power IoT devices autonomously adapt according to the environment to manage on-board and transmission resources efficiently. The devices should observe, learn and then adapt according to changes.

REFERENCES

[1] Faiga Alawad and Frank Alexander Kraemer. 2022. Value of Information in Wireless Sensor Network Applications and the IoT: A Review. *IEEE Sensors Journal* 22, 10 (2022), 9228–9245. https://doi.org/10.1109/jsen.2022.3165946

[2] Iván Alfonso, Kelly Garcés, Harold Castro, and Jordi Cabot. 2021. Self-adaptive architectures in IoT systems: a systematic literature review. *Journal of Internet Services and Applications* 12, 1 (2021), 1–28.

[3] Ali Borji and Laurent Itti. 2012. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence* 35, 1 (2012), 185–207.

[4] Anders Eivind Braten, Frank Alexander Kraemer, and David Palma. 2021. Autonomous IoT Device Management Systems: Structured Review and Generalized Cognitive Model. *IEEE Internet of Things Journal* 8, 6 (2021), 4275–4290. https://doi.org/10.1109/JIOT.2020.3035389

[5] Akram H. Jebril, Aduwati Sali, Alyani Ismail, and Mohd Fadlee A. Rasid. 2018. Overcoming Limitations of LoRa Physical Layer in Image Transmission. *Sensors (Basel, Switzerland)* 18, 10 (2018), 3257. https://doi.org/10.3390/s18103257

[6] Mookeun Ji, Juyeon Yoon, Jeongwoo Choo, Minki Jang, and Anthony Smith. 2019. Lora-based visual monitoring scheme for agriculture iot. In *2019 IEEE sensors applications symposium (SAS)*. IEEE, 1–6.

[7] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Jiacong Fang, imyhxy, Kalen Michael, Lorna, Abhiram V, Diego Montes, Jebastin Nadar, Laughing, tkianai, yxNONG, Piotr Skalski, Zhiqiang Wang, Adam Hogan, Cristi Fati, Lorenzo Mammana, AlexWang1900, Deep Patel, Ding Yiwei, Felix You, Jan Hajek, Laurentiu Diaconu, and Mai Thanh Minh. 2022. *ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference*. https://doi.org/10.5281/zenodo.6222936

[8] Sikandar Zulqarnain Khan, Yannick Le Moullec, and Muhammad Mahtab Alam. 2021. An NB-IoT-Based Edge-of-Things Framework for Energy-Efficient Image Transfer. *Sensors* 21, 17 (2021), 5929.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[10] Louis Moreau. 2022. Announcing Fomo (faster objects, more objects). https://www.edgeimpulse.com/blog/announcing-fomo-faster-objects-more-objects

[11] Henry Muccini, Mohammad Sharaf, and Danny Weyns. 2016. Self-adaptation for cyber-physical systems: a systematic literature review. In *Proceedings of the 11th international symposium on software engineering for adaptive and self-managing systems*. 75–81.

[12] Congduc Pham. 2016. Low-Cost, Low-Power and Long-Range Image Sensor for Visual Surveillance. In *Proceedings of the 2nd Workshop on Experiences in the Design and Implementation of Smart Objects* (New York City, New York) *(SmartObjects '16)*. Association for Computing Machinery, New York, NY, USA, 35–40. https://doi.org/10.1145/2980147.2980156

[13] Peter Schüller, João Paolo Costeira, James Crowley, Jasmin Grosinger, Félix Ingrand, Uwe Köckemann, Alessandro Saffiotti, and Martin Welss. 2022. Composing Complex and Hybrid AI Solutions. https://doi.org/10.48550/ARXIV.2202.12566

[14] Faisal Karim Shaikh and Sherali Zeadally. 2016. Energy harvesting in wireless sensor networks: A comprehensive review. *Renewable and Sustainable Energy Reviews* 55 (2016), 1041–1054.

[15] Anestis Staikopoulos, Venetis Kanakaris, and George A Papakostas. 2020. Image Transmission via LoRa Networks–A Survey. In *2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 150–154.

[16] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (second ed.). The MIT Press.