Multifactorial Uncertainty Assessment for Monitoring Population Abundance using Computer Vision

Emma Beauxis-Aussalet
CWI
Amsterdam, The Netherlands
Email: emalb@cwi.nl

Lynda Hardman
CWI
Amsterdam, The Netherlands
Email: lynda.hardman@cwi.nl

Abstract*—Computer vision enables in-situ monitoring of animal populations at a lower cost and with less ecosystem disturbance than with human observers. However, computer vision uncertainty may not be fully understood by end-users, and the uncertainty assessments performed by technology experts may not fully address end-user needs. This knowledge gap can yield misinterpretations of computer vision data, and trust issues impeding the transfer of valuable technologies. We bridge this gap with a user-centered analysis of the uncertainty issues. Key uncertainty factors, and their interactions, are identified from the perspective of a core task in ecology research and beyond: counting individuals from different classes. We highlight factors for which uncertainty assessment methods are currently unavailable. The remaining uncertainty assessment methods are not interoperable. Hence it is currently difficult to assess the combined results of multiple uncertainty factors, and their impact on end-user counting tasks. We propose a framework for assessing the multifactorial uncertainty propagation along the data processing pipeline. It integrates methods from both computer vision and ecology domains, and aims at supporting the statistical analysis of abundance trends for population monitoring. Our typology of uncertainty factors and our assessment methods were drawn from interviews with marine ecology and computer vision experts, and from prior work for a fish monitoring application. Our findings contribute to enabling scientific research based on computer vision.

I. Introduction

Computer vision technologies can support the study of a variety of animal populations in their natural environment [1], [2], [3], [4], [5]. However, the technical constraints of in-situ video monitoring yield potential errors in the extracted data. Multiple factors are at stake, such as misidentification of animals, camera breakdowns, or encoding errors. Users are aware that computer vision provides uncertain data, yet they may not fully comprehend how it impacts the scientific validity of their data analysis [6]. Users may misinterpret computer vision output, or have uninformed confidence in their interpretation. To address these issues, we investigate the sources of uncertainty in in-situ video monitoring systems, and the means to communicate their impact on end-user tasks.

Our study was performed in the context of the Fish4Knowledge (F4K) project [7], an application of computer vision to the in-situ monitoring of Taiwanese coral reef fish (Fig. 1). It used 9 static cameras, fixed on the seabed at depths around 2 to 3 meters, to continuously record underwater ecosystems for 3 years. Marine ecologists needed to analyze counts of fish for different species, behaviors, time periods or locations. Uncertainty issues were first identified for this context. We discuss them here from a larger perspective including the monitoring of other kinds of individuals. We specify a typology of *uncertainty factors*, i.e., the technological or environmental components yielding uncertainty. We describe the interactions between uncertainty factors, in order to identify i) how uncertainty propagates through the information processing pipeline; and ii) how interoperable assessment methods can describe uncertainty propagation. The goal of our typology is to identify the set of uncertainty measurements that needs to be communicated to end-users, in order to support uncertainty-aware analyses of computer vision data.

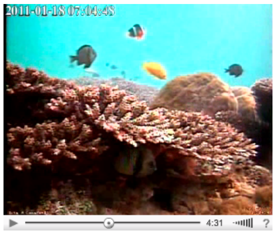

Fig. 1. Example video frame from the F4K system

Existing generic uncertainty typologies [8], [9], [10], [11] can model our typology, which instantiates the case of computer vision systems for in-situ monitoring of animal populations. Walker et al. (2003) caution against *"framing problems such that the context fits the tacit values of the experts and/or fits the tools, which experts can use to provide a solution to the problem"* [8]. *Locations* of uncertainty within the video monitoring system [8] may lie beyond the set of uncertainty factors addressed by a single domain of expertise. We thus consulted multidisciplinary experts with different specialties within the marine ecology and computer vision domains, and considered potential sources of uncertainty at each step of the information processing pipeline. Consistent with Pang et al. (1997), our uncertainty factors can be mapped to a 3-step pipeline of data collection (i.e., the in-situ deployment of the system), data processing (i.e., the computer vision algorithms) and data interpretation (i.e., the analysis of computer vision outputs) [10]. Correa et al. (2009) further discuss uncertainty propagation and aggregation along the processing steps [11]. Our framework for specifying uncertainty propagation integrates methods from both computer vision and ecology domains. From this framework, we elicit a set of key uncertainty measurements supporting uncertainty-aware analyses of population abundance, and identify unaddressed problems requiring further research. Our approach contributes to data science by making video monitoring systems meet the scientific requirements with i) transparent uncertainty factors; and ii) methods to take uncertainty into account in end-user tasks.

II. Background Literature

Computer vision uncertainty - Computer vision uncertainty is mainly evaluated using groundtruth datasets, which consist of manually classified images (e.g., examples of animals for each species). The groundtruth is split into two distinct sets: one for *training* the algorithms, one for *testing* their results. From the *training set*, algorithms construct a model of the items to classify. It describes the features that are representative of items' visual appearance (e.g., shape, color, texture). With the *test set*, the computer vision results are compared to the manual classifications. The primary uncertainty assessment is the number of misclassifications: items' *true class* is known from the manual classification, and the algorithms' *output class* may give correct classifications (True Positives *TP*, True Negatives *TN*) and misclassifications (False Negatives *FN*, False Positives *FP*). Errors are encoded in confusion matrices (Fig. 2) from which a variety of uncertainty metrics can be derived. Derived metrics are mainly proportions of TP, TN, FN and FP relative to total numbers of Positive or Negative items, e.g., FP Rate, Precision, Recall [12], [13].

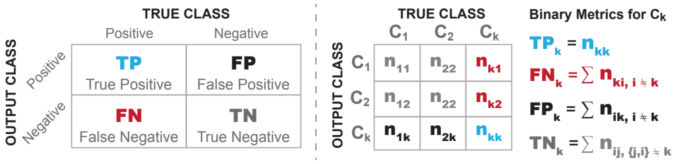

Fig. 2. Typical confusion matrices for binary problems (left), e.g., detection of individuals (TP, FN) and other objects (TN, FP), and multiclass problems (right), e.g., recognition of multiple classes of species or behaviors.

Uncertainty can also be measured as the similarity between an item and the class model constructed from the training set. *Similarity measures* can indicate that an item's appearance is, e.g., 71% similar to the model. The lower the similarity, the higher the chance of misclassification. Similarity measures can be computed using various methods depending on the application context [14]. They can be available for any classified item, including items for which no groundtruth classification is available. Thus they can indicate uncertainty when the groundtruth's true class is unknown. They are often used as a threshold for discarding low-similarity items likely to be False Positives. They can also be used to infer error probabilities depending on similarity measures, and correct counts of individuals accordingly, e.g., Boom *et al.* [15]. However, their method requires similar class proportions in test sets and end-results. This cannot be ensured in population monitoring as the relative abundance of species and behaviors may vary over time and location (e.g., due to migration or reproduction cycles).

After detecting objects occurring in each video frame, computer vision systems apply tracking algorithms to identify individuals and their trajectories across frames. Tracking errors can occur between each pair of frames, and impact the resulting counts of items per class: i) one single object may not be linked from one frame to the other, and considered as 2 distinct objects (thus over-estimating counts of items); ii) 2 distinct objects may be erroneously linked and considered as one single object (thus under-estimating counts of items); or iii) the trajectories of adjacent objects may be confused with one another (thus increasing the chances of misclassifying species and behaviors). Evaluations of tracking errors compare groundtruth trajectories with algorithm outputs. Palazzo et al. (2012) propose 3 evaluation metrics reflecting i) the tracking errors at each pair of frames (*correct decision rate*); ii) the consistency of single trajectories which may contain images from different objects (*trajectory matching*); and iii) the resulting counts of objects per class (*correct counting rate* $\frac{TP}{TP+FN}$, i.e., Recall) [16].

For computer vision systems applying a sequence of algorithms (e.g., Fig. 3), uncertainty propagates over the information processing pipeline. Each classifier's uncertainty depends on its internal parameters (e.g., for processing low-level image features), *training set* and resulting class models, and *input set* (e.g., a set of images to classify). Labels from previously applied classifiers (e.g., bounding box with identified individuals and/or species) may be included in input sets as a categorical attribute, with corresponding *similarity measures* as a numerical attribute. In practice, misclassifications are measured for the end-results of the whole processing pipeline, or worse, for each classifier separately and potentially with heterogeneous *test sets*. Crosetto et al. (2001) provide foundations for evaluating uncertainty considering i) the algorithms' pipeline as a black box; ii) algorithms' parameters (threshold on *similarity measures*); and iii) input sets of numerical attributes with systematic errors (bias) and stochastic errors (noise) [17]. Zhu et al. (2004) study uncertainty propagation w.r.t. i) training sets with stochastic errors in one or several numerical and categorical attributes; ii) training sets with stochastic errors in the groundtruth class; iii) input sets with stochastic errors in one or several numerical and categorical attributes [18]. Senge et al. (2014) investigate uncertainty propagation for ensemble classifiers, i.e., a pipeline of binary classifiers providing multiclass classification as end-results [19]. No prior work was found to address the uncertainty propagation over pipelines including binary and multiclass classifiers, and tracking algorithms; and with systematic errors in the categorical attributes of input sets. The latter needs to be considered to assess uncertainty propagation through imperfect classifiers. For instance, in Fig. 3, *Species Recognition* may systematically confuse specific species, and *Behavior Recognition* may process input sets with systematic errors in the species attribute.

In-situ video monitoring of populations - Automating the recognition of animals is already challenging in controlled environments [1], and more so in-situ. Video technologies have been applied to ecosystem monitoring with a variety of settings, e.g., for marine ecosystems: fields of view, lighting [20], [21], baited or unbaited, and static or diver-operated cameras [22], [23], manual or automated analysis [24], color filters, lenses, and parameters of video analysis algorithms [20]. These settings impact the uncertainty of computer vision results, e.g., error magnitude can vary over lighting or field of view. Error magnitude can also vary over species and species

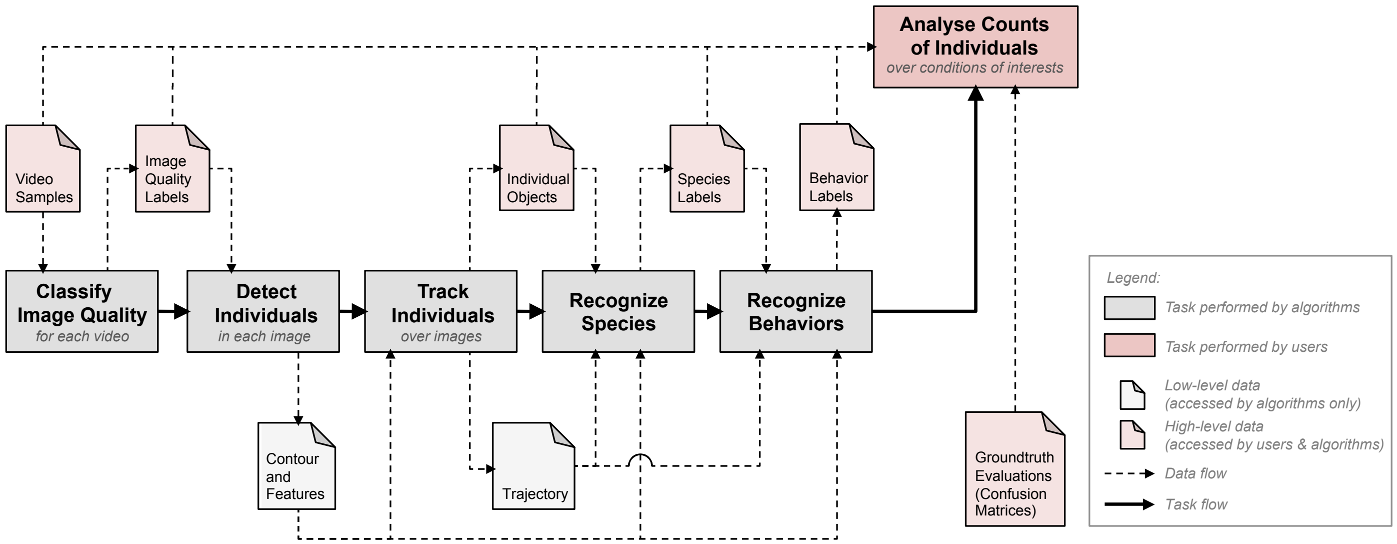

Fig. 3. Typical data processing pipeline for the video monitoring of animal populations (BPMN notation)

can be systematically confused, thus biasing the data [1]. Ecologists usually deal with uncertainty using sampling theory and by applying repeated measurements and statistics (e.g., variance, ANOVA) [25], [26]. No well established guidelines were found to integrate classification errors from computer vision into standard statistical frameworks used for monitoring population abundance. However, computer vision uncertainty is likely to be accepted by ecologists [27] as i) coping with uncertainty is part of their current practices; ii) computer vision benefits overcome the technical issues; and iii) the technical issues can be improved [1], [28].

III. Context of Application

Typical use case for population monitoring - User information needs were investigated for the Fish4Knowledge fish monitoring system [27], [29]. Ecologists needed to study the population dynamics, i.e., the evolution of population sizes, and underlying phenomena such as reproduction, migration and trophic systems (food chains). The primary information need is to obtain counts of individuals over 5 dimensions of interest: time period, location, species, behavior and body size (Table II). For technical reasons, the project was not able to deliver information on body size and behavior. However, uncertainty issues with behavior recognition are included in this paper and the complete workflow addressing these information needs is shown in Fig. 3. Body size is excluded as it requires other types of technology, such as stereoscopic vision. Another key information need is to assess the computer vision uncertainty and how it impacts the scientific validity of data analysis and interpretation. We address this need by complementing prior studies in [6], [27], [30] with 4 contributions: i) a comparison of the potential biases of computer vision to that of other data collection methods (Table III); ii) a refined specification of 12 key uncertainty factors (Table IV) and their interactions (Fig. 4); iii) novel methods for measuring errors and biases (equations 2-8); and iv) a comprehensive framework of uncertainty measurements (Fig. 6).

Typical framework of computer vision algorithms - Video monitoring systems may apply different algorithms depending on use cases, using different classification techniques (e.g., SVM, Bayes, GMM) and low-level feature extraction methods (e.g., Fourier descriptor, Gabor filter, Histogram of Oriented Gradients, Moment Invariants). However, for the vast majority of use cases, algorithms perform 3 high-level tasks: binary classification (e.g., *Detect Individuals*), multiclass classification (e.g., *Recognize Species*) and tracking. We focus on a typical computer vision framework with algorithms performing these 3 tasks (Fig. 3). Our scope of algorithms excludes low-level sub-processing algorithms that are not directly related to our counting task. For instance, algorithms which *Detect Individuals* use binary algorithms for classifying each pixel as being within or outside an object contour (segmentation). Imperfect segmentation influences uncertainty of higher-level algorithms, but measuring segmentation errors does not contribute to estimating errors in counts of individuals. However, estimating sizes of areas (e.g., land coverage estimated from satellite images) requires estimating segmentation errors. Our method for estimating unbiased population abundances (e.g., equation 8) is, however, also applicable to estimating areas, as counting pixels per class is conceptually similar to counting individuals per class.

The use cases captured by our framework in Fig. 3 may have chosen alternative implementation strategies. For instance, *Recognize Species* may be performed before *Track Individuals*, and species labels may be used as input features for *Track Individuals*. As shown in Section V, this alternative does not impact the uncertainty assessment methods we propose. However, our methods rely on 2 assumptions on the framework of computer vision algorithms: i) Video samples are of equal duration, e.g., in Fish4knowledge, continuous video streams are split into 10-minute samples, which simplifies the uncertainty assessment; ii) Image quality is assessed for each video sample, and classified into several categories. Image quality could be measured with continuous values (e.g., blur score), or be measured within each image (segmentation). Opportunities of such approaches are worth being investigated in future work. For the sake of simplicity, we assume that the identification of image quality (e.g., *Recognize Image Quality* in Fig. 3) yields no classification error.

TABLE I. KEY UNCERTAINTY FACTORS IN COMPUTER VISION SYSTEMS FOR POPULATION MONITORING.

Factor	*Description*
Uncertainty due to computer vision algorithms	
Groundtruth Quality	Groundtruth items may be scarce, represent the wrong animals, odd animal appearances (i.e., odd feature distributions).
Object Detection Errors	Some individuals may be undetected, and other objects may be detected as individuals of interest.
Tracking Errors	Trajectories of individuals tracked over video frames may be split, merged or intertwined.
Species Recognition Errors	Some species may not be recognized, or confused with another.
Behavior Recognition Errors	Some behaviors may not be recognized, or confused with another.
Uncertainty due to in-situ system deployment	
Field of View	Cameras may observe heterogeneous ecosystems, and over- or under-represent species, behaviors or objects features. Fields of view may be partially or totally occluded, and shift from their intended position.
Fragmentary Processing	Some videos may be yet unprocessed, missing, or unusable (e.g., encoding errors).
Duplicated Individuals	Individuals moving back and forth are repeatedly recorded. Rates of duplication vary among species behaviors and *Fields of view*.
Sampling Coverage	The numbers of video samples may not suffice for software outputs to be statistically representative.
Uncertainty due to both computer vision algorithms and in-situ system deployment	
Image Quality	Lighting, water turbidity, contrast, resolution or fuzziness may impact the magnitude of computer vision errors.
Noise and Biases	Computer Vision errors may be random (noise) or systematic (bias). Biases may emerge from a combinaison of factors (*Image Quality, Field of View, Duplicated Individuals, Object Detection Errors, Species & Behavior Recognition Errors*).
Uncertainty in Specific Output	Uncertainty in specific sets of Computer Vision output may be extrapolated from errors measured in test conditions, taking into account the specific characteristics of output sets (e.g., fewer low-quality images).

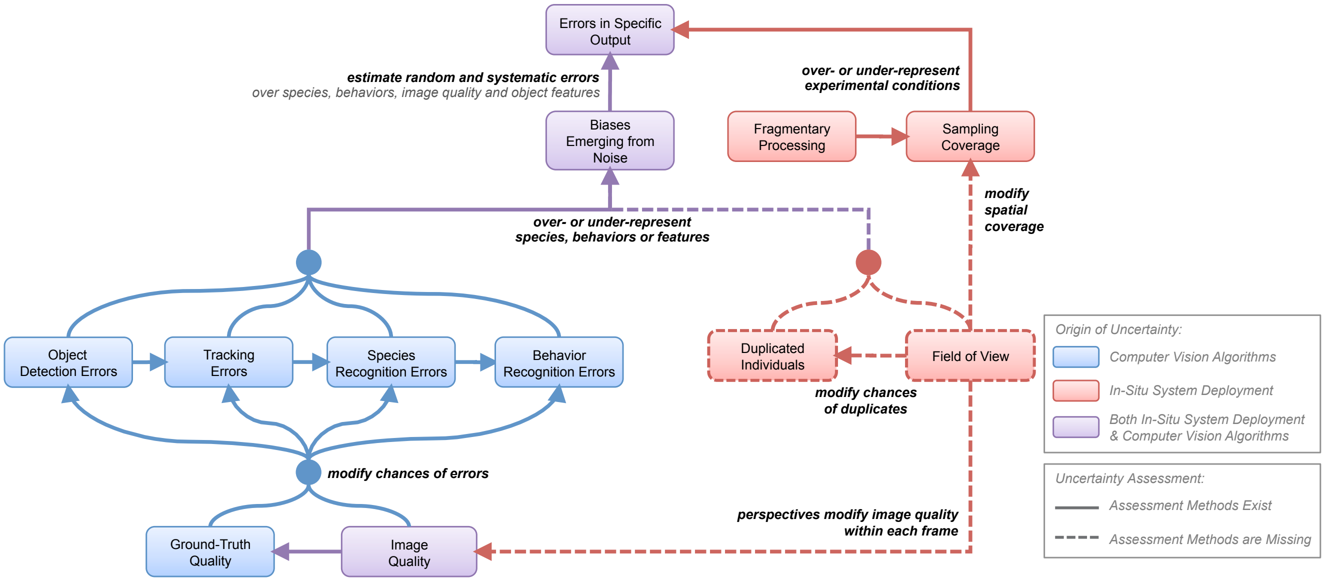
Errors in Specific Output
estimate random and systematic errors
over species, behaviors, image quality and object features
Biases Emerging from Noise
over- or under-represent experimental conditions
Fragmentary Processing
Sampling Coverage
modify spatial coverage
over- or under-represent species, behaviors or features
Object Detection Errors
Tracking Errors
Species Recognition Errors
Behavior Recognition Errors
Duplicated Individuals
Field of View
modify chances of duplicates
Origin of Uncertainty:
Computer Vision Algorithms
In-Situ System Deployment
Both In-Situ System Deployment & Computer Vision Algorithms
modify chances of errors
Uncertainty Assessment:
Assessment Methods Exist
Assessment Methods are Missing
perspectives modify image quality within each frame
Ground-Truth Quality
Image Quality

Fig. 4. Interactions of Uncertainty Factors

IV. UNCERTAINTY FACTORS

Considering the entire population monitoring system, potential errors and biases are not only due to computer vision software (*data processing* in [10]). Uncertainty is also introduced throughout the in-situ deployment of the system (*data collection*). For example, some cameras may receive less light, yielding poor image quality and more computer vision errors. At the *data interpretation* stage, ecologists need to consider uncertainty factors from both computer vision and system deployment, as their interactions yield potential errors and biases in the counts of individuals. Ecologists also need to consider the characteristics of the data subsets under analysis, as these impact the level of uncertainty (e.g., more low-quality images compared to the overall dataset increases uncertainty). Overall, 12 key uncertainty factors were identified along the information processing pipeline, as summarized in Table IV.

Uncertainty due to computer vision algorithms - Computer vision algorithms use groundtruth training sets to learn to detect individuals, species or behaviors, but also to track individuals and to detect image quality. Groundtruth is typically manually annotated by experts, but is often crowdsourced by non-experts [31]. **Groundtruth Quality** is essential to control the errors in computer vision results. Scarcity, unrepresentative image quality or annotation errors in groundtruth sets may yield error-prone computer vision software. **Image Quality** impacts the appearance of objects, and thus the consistency of visual features extracted by computer vision algorithms and used to recognize animals, species and behaviors. The Fish4Knowledge system was designed to classify image quality, and apply class-specific preprocessing parameters to extract consistent fish features regardless of the original image quality. To infer the impact of image quality on computer vision errors, groundtruth evaluations must be performed using samples from each image quality class. As the errors of each classifier impact that of the next classifiers, groundtruth and image quality are important components of uncertainty propagation.

Computer vision may result in 4 types of high-level errors. **Object Detection Errors** concern the detection of individuals in each video frame, i.e., undetected individuals (FN) and other objects identified as individuals of interest (FP). **Tracking**

TABLE II. INFORMATION NEEDS & DATA COLLECTION METHODS [27]

	Fish Counts	Species Recognition	Behavior Recognition	Fish Size
Research Topic				
Population Dynamics	mandatory	mandatory	optional	important
Trophic Systems	mandatory	mandatory	important	important
Reproduction	mandatory	mandatory	important	important
Migration	mandatory	mandatory	optional	optional
Data Collection Method				
Scientific Fishery	+	+/++[1]	-	+
Commercial Fishery	+	+	-	+
Diving Observation	+	+	++	+
Manual Image Analysis	+	+	+	-/+ [2]
Computer Vision	+	+	-/+[3]	-/+ [2]

The signs indicate whether data collection methods: **-** *cannot supply the information,* **+** *can supply the information,* **++** *can supply the most precise information.*

[1] *Fish dissection performed after scientific fishing is the most accurate. technique for recognizing coral reef species that are visually similar.*

[2] *Possible with stereoscopic vision, or calibrated distance camera-background.*

[3] *The state-of-the-art does not fully address the wide scope of fish behavior variety.*

TABLE III. BIASES WITH SPECIES & DATA COLLECTION METHODS

Type of Species	Scientific Fishery	Commercial Fishery	Diving Obs.	Manual Img Analysis	Computer Vision
Benthic	-[1]	-[1]	=	=	=[2]
Sedentary	-	-	=	=	=/+[3]
Schooling	=	=	-/+	-/+	-/+[3,4]
Small	-/=[4]	-/=[5]	-/=[6]	-/=[6]	-/=[6]
Shy	-	-	-/=[7]	-/=[8]	-/=[8]
Cryptic	-	-	=	-	-
Look-alike	=	=	-/+	-/+	-/+
Rare	=	-	=	=	-/=[9]
Nocturnal	=	=	-	-/=	-/=[10]
Carni- or Herbivorous	-/=/+[11]	=	=	-/=/+[11,12]	-/=/+[11,12]

The signs indicate whether parts of ecosystems are likely to be **+** *over-represented,* **=** *neither under- nor over-represented,* **-** *under-represented.*

[1] *Considering that the destructive use of trawl nets is not an option.*

[2] *If cameras' field of view target the seafloor.*

[3] *Species often swimming in and out cameras' field of view are over-estimated.*

[4] *Fish in groups occlude each other and are under-estimated.*

[5] *Large granularity of nets' and fish traps' mesh can let small fish slip through.*

[6] *Small fish may not be visually detectable from a large distance.*

[7] *Cloaking procedures can allow the observation of shy fish.*

[8] *With handheld cameras, some species flee from divers.*

[9] *Recognizing all rare species may not be possible due to lack of ground-truth.*

[10] *Possible with night vision cameras.*

[11] *Baits, if used, can attract either herbivorous or carnivorous species.*

[12] *Cameras' field of view may overestimate some species and underestimate others.*

Errors concern the misidentification of individual trajectories across multiple frames, i.e., split or merge individual trajectories, or intertwined trajectories of different individuals. **Species Recognition Errors** are individuals recognized as a species they do not actually belong to. **Behavior Recognition Errors** are individuals recognized with a behavior they are not actually exhibiting.

Uncertainty due to in-situ system deployment - This source of uncertainty is usually not in the scope of evaluations performed in computer vision research. Evaluations of computer vision algorithms are intended to be valid for most applications of the algorithms, and are abstracted from case-specific application conditions. However, errors and biases in computer vision outputs can be significantly influenced by i) time-varying environmental conditions (e.g., lighting, turbidity, biofouling) or camera features (e.g., lens, resolution) that lower **Image Quality**; ii) the placement of cameras, i.e., the **Field of View** can target specific habitats and under-represent species living in other habitats, over-represent animal behaviors occurring in these habitats, modify the chances of **Duplicated Individuals** (e.g., targeting a feeding zone may increase the number of individuals moving back and forth, thus in and out the field of view), or modify the chances of obtaining low *Image Quality* (e.g., in shade- or turbidity-prone locations); iii) the numbers of cameras which may not provide sufficient **Sampling Coverage**; and iv) computational issues with the servers executing the computer vision algorithms, which can yield **Fragmentary Processing** (e.g., missing videos).

Synthesizing uncertainty in specific outputs - Ecologists are concerned with differentiating stochastic errors (random noise) from systematic errors (bias). **Noise and Biases** arise from a combinaison of factors yielding counts of individuals that are systematically lower or higher than their true values (i.e., compared to counts from groundtruth sets). The levels of noise and bias observed for groundtruth sets may differ from that of specific subsets of computer vision output. Hence for deriving the **Uncertainty in Specific Outputs** from the groundtruth evaluations, ecologists must account for the specific characteristics of data subsets. They need to assess i) the proportion of *Image Quality* in groundtruth and output sets, to infer the magnitude of errors in output sets given the error magnitude measured for each image quality; ii) how *Fields of View* impact the chances of *Duplicated Individuals* and the completeness of *Sampling Coverage*, as these potentially under- or over-estimate some species or behaviors (e.g., Table III).

Interactions of uncertainty factors - The uncertainty factors interact with each other, yielding a complex scheme of uncertainty propagation (Fig. 4). Computer vision algorithms are impacted by the errors of the algorithms previously applied. *Object Detection Errors* impact *Tracking Errors* as missing individuals (FN) and other objects (FP) yield erroneous interpretations of trajectories. *Species Recognition Errors* are impacted by both *Object Detection* and *Tracking Errors*, as FP objects may be attributed a species, and species recognition suffers from intertwined trajectories merging individuals from different species. *Behavior Recognition Errors* are impacted by *Species Recognition Errors* as behavior features are species-specific (e.g., one speed indicates hunting behavior for one species, but is a neutral movement for another).

The *Field of View* impacts the kind of ecosystems observed by each camera. It also impacts the chances of *Duplicated Individuals*, e.g., observing coral heads is more likely to yield overestimation of sedentary species than observing the open sea. Hence the *Field of View* can over-estimate local species and behaviors, and under-estimate others, thus influencing the *Noise and Biases*. The depth of *Field of View* impacts the size of the monitored areas, hence the *Sampling Coverage*. It further impacts *Image Quality* as resolution and fuzziness are poorer for the background than the foreground. In this case, approaches classifying image quality within each frame (segmentation) may be of interest. *Image Quality* is further impacted by the *Field of View* as a camera may be placed in area where low-lighting, turbidity or bio-fouling are more likely to occur. Different classes of *Image Quality* can yield different levels of *Object Detection, Species Recognition* and *Behavior Recognition Errors*, and thus potential *Noise and Biases*. Hence *Groundtruth Quality* depends on how image samples are representative of the range of possible *Image Quality*. The groundtruth samples need to represent the possible appearances of individuals (e.g., different angles), and contain multiple samples of appearances to account for the statistical variations of the low-level image features (e.g., variability of color rendering or contours). Finally, the initial *Sampling Coverage* of the cameras deployed over ecosystems can be reduced by the *Fragmentary Processing* of the videos, i.e., due to unprocessed or missing videos.

V. Uncertainty Assessment

For computer vision experts, the primary uncertainty assessment methods are groundtruth evaluations, performed for each algorithm or for the whole pipeline of algorithms (a black box). End-users who are not familiar with computer vision are likely to encounter difficulties in understanding groundtruth evaluations, confusion matrices and their technical concepts [27]. Derived uncertainty metrics (e.g., Precision, Recall) may be misunderstood, and do not fully address all uncertainty factors. When integrating computer vision data into their scientific research, ecologists may not know i) how misclassifications impact the counts of individuals in end-results; and ii) how to combine their statistical methods with measurements derived from confusion matrices. We address these issues with assessment methods for uncertainties due to computer vision algorithms and in-situ system deployment, the related uncertainty propagation, and the resulting impact on counts of individuals.

Assessing computer vision algorithms - Confusion matrices need to be read both column- and row-wise, which is tedious and error prone. Considering the Class C_k in Fig. 2, if read row-wise the matrix indicates FP added to C_k. If read column-wise, it indicates FN lost by C_k. Memorizing all cell values, and their meaning, requires an important cognitive effort. Users may forget cell values, or may read only columns or rows hence omitting type I or II errors.

Multiclass confusion matrices are usually synthesized by summing errors for each class (Fig. 2 extreme right). However, it is no longer possible to distinguish which classes are likely to be confused with one another, e.g., summed FP do not indicate the original true species of the misclassified individuals. Users need this information to identify biases induced by *Species* or *Behavior Recognition Errors*, and hence, counts of individuals that are not representative of the actual population dynamics. For instance, an increase of one species implies an increase of its FN. A proportion of its individuals are likely to be systematically attributed to other species. This can induce a deceiving increase of another species population, especially for species of much lower abundance. Hence, users need to inspect the errors between pairs of species, rather than only the synthesis of FP and FN summed for all species.

Derived metrics such as *FP Rate* ($\frac{FP}{FP+TN}$), *FN Rate* ($\frac{FN}{FN+TP}$) and *Accuracy* ($\frac{TP+TN}{TP+TN+FP+FN}$) are complex and convey specific types of errors. Non-expert users may misinterpret them. For instance, when recognising species S_k, TN are individuals correctly discarded from the S_k population. TN are summed for all species other than S_k, and are usually of a much higher magnitude than the TP, FP and FN for S_k. High numbers of TN yield low *FP Rates*, which may conceal important numbers of FP and FN. Uncertainty assessments commonly performed by computer vision experts use pairs of advanced metrics (e.g., FP and FN Rates in ROC curves [13]). Such sets of metrics are likely to overwhelm and mislead users who are not familiar with computer vision. Moreover, advanced metrics no longer indicate the number of items in the groundtruth, and thus possible groundtruth scarcity which is an important aspect of *Groundtruth Quality*.

We elicited a set of uncertainty metrics that address uncertainty factors due to computer vision algorithms, while limiting misunderstandings and cognitive load. We first retain the design choice to omit TN from [32]. They are not contained in end-results and are not informative for users. Further, [27] shows that understanding the concepts of TP, FN and FP is already likely to overwhelm users. To limit cognitive load and misunderstandings, we avoid advanced metrics and primarily provide the numbers of TP, FN and FP rather than error rates. Hence we convey the numbers of groundtruth items (which are usually proportional in training and test sets), an important aspect of *Groundtruth Quality* albeit abstracted in traditional ROC or Precision/Recall curves.

Numbers of groundtruth items can greatly vary amongst classes of species, behaviors or image quality, e.g., scarcity for some classes, or abundance of others. In these cases, classification errors can be difficult to compare across classes. Hence we convey magnitudes of errors either as: i) numbers of groundtruth items (default); or ii) proportional measure of errors (on-demand). We retained the novel error rate formula (1) from [15], [33] as it can convey the impact of misclassifications on end-user counting tasks. Users need to estimate the true counts (TP+FN) from the output counts (TP+FP), e.g., the proportions of FP and FN w.r.t. the output counts they are provided with. Precision ($\frac{TP}{TP+FP}$) and Recall ($\frac{TP}{TP+FN}$) target such estimations. However, the use of FP is problematic since FP and FN are not independent: FN for one class are FP for others. Hence, the number of FP observed for one class varies depending on the number of groundtruth items for other classes. In ecosystems, the relative abundance of species and behaviors may greatly vary over time or location, and may not be similar to that of the groundtruth set. Therefore FP are excluded from the measure of errors. FN transferred from Class C_a to C_b (i.e., $FN_{a \to b}$) are given proportionally to the TP of their true class C_a, using the equation (1). Although atypical, the error ratio supports the estimation of errors in counts of individuals, while accounting for potential changes in class proportions. Further, it supports visualizations where both FP and FN are represented with bars aligned horizontally (Fig. 5), thus facilitating the comparisons of type I and II errors over classes.

$$Pairwise\ Error\ Ratio\ C_a \to C_b = \frac{FN_{a \to b}}{TP_a} \quad (1)$$

Equation 1. $Pairwise\ Error\ Ratio\ C_a \to C_b$ [33] is the ratio of individuals belonging to class A (C_a) erroneously attributed to class B (C_b). $FN_{a \to b}$ is the number of groundtruth items attributed to C_b while truly belonging to C_a. TP_a is the total number of TP for C_a. Note that $FN_{a \to b}$ is different from $FN_{b \to a}$ and $Pairwise\ Error\ Ratio\ C_a \to C_b$ is different from $Pairwise\ Error\ Ratio\ C_b \to C_a$.

This approach was presented to ecology, computer vision and visualization experts [30], [32], [34] and received positive feedback for its simplicity and understandability. It supports the assessment of uncertainty due to computer vision algorithms (*Image Quality, Groundtruth Quality, Animal Detection, Species Recognition* and *Behavior Recognition Errors*), and conveys a simplified but complete representation of groundtruth evaluations. However, other aspects of *Image* and *Groundtruth Quality* are not included. But these are of interest mostly for technology providers rather than end-users, or are of little interest for assessing the uncertainty in counts of individuals. Other aspects of image quality include, for instance, how many classes of *Image Quality* are relevant, or how preprocessing achieves to provide consistent feature descriptors regardless of the original image quality. Or for *Groundtruth Quality*, how many persons manually classified each items, what was their expertise, and did their classifications differ (e.g., Cohen's Kappa).

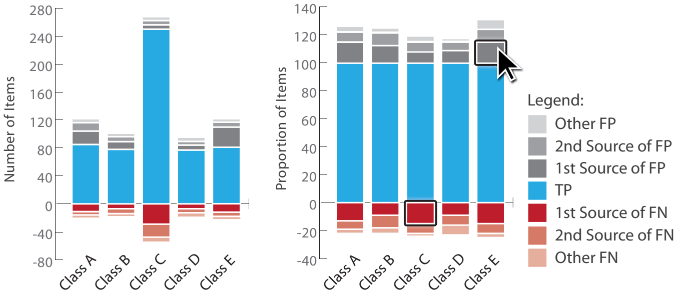

Fig. 5. Visualizations for the retained uncertainty assessment approach [34]

Assessing uncertainty due to in-situ system deployment - The state-of-the-art does not offer well-established methods for handling these uncertainty factors, as technical experts focus on generic uncertainty assessment abstracted from application-specific conditions. Future work needs to develop measures of *Duplicated Individuals* depending on species, behaviors and *Fields of View*. For example, classes of fields of view may be identified, and rates of duplicates be estimated for each species, e.g., species S_i observed from field of view V_j are impacted by $Rate\ of\ Duplicate_{i,j} = \frac{True\ Count}{Output\ Count}$. Species living in groups may be under-estimated due to occlusions, and the related uncertainty may be assessed with such rate of duplicates. With monitoring systems using fixed cameras, fields of view may gradually vary over time (e.g., with maintenance, lens cleaning, typhoons or other incidents). Assessment methods are missing for this difficult problem.

The impact of *Sampling Coverage* and *Fragmentary Processing* can be assessed using sampling theory. Ecologists need to take into account the number of video samples from which computer vision results are drawn, since it influences the statistical representativity of the observed counts of individuals. For instance, counts observed from a few videos may not be representative of the actual population abundance. Further, if counts are drawn from different numbers of video samples, the more videos the more individuals, hence comparison is biased. Therefore users need evaluations of sampling size (e.g., the numbers of equal-duration videos), and a comparable measure of abundance for end-results drawn from different sampling sizes. As we use equal-duration videos, the primary metric for sampling size is the number of video samples from which end-results are extracted. To analyze sets of end-results extracted from varying numbers of video samples, averaging the counts of individuals per video as in (2) offers a comparable metric of abundance. $Mean\ Count\ per\ Video$ must be analysed with care as 3 problems may arise: 1) Video duration must be identical over samples; 2) Combining data from several cameras and time periods involves subtleties; and 3) the computation of variance face issues with *Sampling Coverage* and *Fragmentary Processing*.

$$Mean\ Count\ per\ Video = \frac{Number\ of\ Individuals}{Number\ of\ Videos} \tag{2}$$

Equation 2. Measure of population abundance for comparing counts of individuals drawn from varying numbers of videos.

1) *Heterogeneous video durations* bias the $Mean\ Counts\ per\ Video$ (2), as longer videos contain more individuals than shorter ones. Equations could be modified to handle videos of unequal durations, e.g., by computing counts per video as $\frac{Number\ of\ Individuals\ in\ Video\ i}{Duration\ of\ Video\ i}$.

2) *Combining data over different cameras and time periods* is subtle. It may seem trivial to compute population abundance over different cameras as $\frac{Total\ Individuals\ for\ All\ Cameras}{Total\ Videos\ for\ All\ Cameras}$, but it is incorrect. For instance, with 2 video samples recorded simultaneously from two cameras (with different fields of view), and yielding n_1 and n_2 counts of individuals, the total abundance is $n_1 + n_2$ rather than $\frac{n_1+n_2}{2}$. This is because the 2 videos record the same time period, and the overall abundance is the sum of individuals occurring at the different locations. The measure in (3) is the correct alternative. It must be interpreted as the mean count of individuals for the time unit represented by the video duration (e.g., mean abundance per 10-min for the Fish4Knowledge project). Note that this approach is not appropriate if cameras observe overlapping fields of view.

$$Mean\ Count\ over\ Cameras = \sum_{j=1}^{N_c} MCV_j \tag{3}$$

Equation 3. Measure of abundance over several cameras. N_c is the number of cameras, and MCV_j is the $Mean\ Count\ per\ Video$, i.e., the result of equation (2), for camera j.

3) *The variance of population abundance* may be difficult if video samples are missing. For mean counts from a single camera (2), equation (4) gives the variance over videos. Mean counts over several cameras (3) is a sum of random variables, which variance is given by equation (5). Measuring the variance of (3) assumes that video samples are available for all cameras. For instance, if one cameras provides a video sample for the time period t (e.g., 08:00 to 08:10 on Jan. 1st 2012) then all cameras must provide a video sample for that period. Otherwise neither the covariance (6) nor the variance of $Mean\ Count\ over\ Cameras$ (5) can be computed. For such issues due to *Fragmentary Processing* or heterogeneous *Sampling Coverage*, alternative methods exist [35] and must be chosen depending on application requirements.

$$Single\ Camera\ Variance = \frac{1}{N_v} \sum_{i=1}^{N_v} (MCV - N_i)^2 \tag{4}$$

Equation 4. Measure of variance in population abundance measured at a single camera. N_v is the number of video samples, N_i is the count of individuals in the i-th video, and MCV is the $Mean\ Count\ per\ Video$, i.e., the result of equation (2).

$$Variance\ over\ Cameras = \sum_{j=1}^{N_c} \sum_{k=1}^{N_c} Cov(MCV_j, MCV_k) \quad (5a)$$

$$= \sum_{j=1}^{N_c} Var(MCV_j) + \sum_{j=1}^{N_c} \sum_{k>j}^{N_c} Cov(MCV_j, MCV_k) \quad (5b)$$

Equation 5. Measure of variance in population abundance measured over several cameras. N_c is the number of cameras, MCV_j is the $Mean\ Count\ per\ Video$, i.e., equation (2), at camera j, $Var(MCV_j)$ is the result of equation (5) for Camera j, and $Cov()$ is the covariance from equation (6).

$$Cov(MCV_j, MCV_k) = \frac{1}{N_t} \sum_{t=1}^{N_t} (MCV_j - N_{j,t})(MCV_k - N_{k,t}) \quad (6)$$

Equation 6. Measure of covariance used in equation (5). N_t is the number of time periods of the duration of a video sample (e.g., 10min for Fish4Knowledge) for which video samples are available for all N_c cameras, $N_{j,t}$ is the number of individuals at camera j during time period t, and MCV_j is the $Mean\ Count\ per\ Video$, i.e., the result of equation (2), for camera j.

Assessing uncertainty in output counts - Although there is a variety of factors and interactions between them, ecologists seek to synthesise the overall impact of the various uncertainty factors as two types of effect: noise (i.e., random errors yielding count variance) and biases (i.e., systematic errors yielding under- or over-estimated counts). Random errors are commonly measured using metrics of mean and variance (equations 2-6). These metrics are well-established bases for the statistical analysis of populations [36]. Significant differences of means and variances may be observed due to technical features, e.g., image quality, rather than natural phenomena. If so, the technical features potentially introduce biases. As no well-established methods are available for evaluating biases due to *Field of View* and *Duplicated Individuals*, the rest of the discussion focuses on identifying biases introduced by computer vision algorithms.

The groundtruth evaluation measuring *Object Detection, Species Recognition* and *Behavior Recognition Errors* can also support the evaluation of biases due to *Image Quality* and *look-alike* species or behaviors. If error measurements (i.e., equation 1) are significantly different amongst *Image Quality*, it indicates potential biases in the counts of individuals. Counts of individuals from a specific image quality can be artificially over- or under-estimated, compared to counts from another image quality. If error rates are of the same magnitude for all image qualities, it indicates a general level of noise unlikely to yield biases, even if error rates are high. Counts drawn from different image qualities, and having similar magnitudes of errors, are potentially over- or under-estimated in the same way, and hence, are comparable.

In contrast, high error rates for *Species* or *Behavior Recognition Errors* indicate potential biases between look-alike species or behaviors, even if error rates are of the same magnitude for all classes. As a class abundance varies over time (e.g., migration or reproduction periods), the magnitude of its FN vary accordingly. These FN are attributed to other classes whose abundances may also vary. Hence the counts of individuals may be artificially correlated, and may not represent the actual trends in the observed ecosystem. The Pairwise Error Ratio (1) and visualizations in Fig. 5 support the identification of such biases with look-alike species and behaviors. We further propose a novel method to correct for potential biases, and estimate the correct counts of items. Our method, called PERLE for Pairwise Error Rates and Linear Equations, is given by equation (8), using the notation in Table IV. Variables noted with the prime symbol designate counts in end-results, e.g., $n'_{.i}$ is the output count for a given class, which is known *a priori*, and $n'_{i.}$ is the true count, which is unknown *a priori*. Variables noted without the prime symbol designate counts for the groundtruth dataset, e.g., $n_{.i}$ is the output count, $n_{i.}$ is the true count, and both are known *a priori*. Our method relies on the assumption that error rates are equivalent in groundtruth and end-results. Given $\frac{n'_{ki}}{n'_{k.}} = \frac{n_{ki}}{n_{k.}}$ thus $n'_{ki} = n'_{k.} \frac{n_{ki}}{n_{k.}}$ and $n'_{.i} = \sum n'_{ki}$ we can construct the linear system (7). Its unknown variables are the true counts $n'_{i.}$. They can be derived by solving the linear system, i.e., by inverting the error rate matrix in (8) and multiplying the result by the vector of output counts. A numerical example of PERLE results is provided in Table V. The applicability and limitations of PERLE needs to be investigated in future work with real and synthetic datasets.

$$\left\{\begin{array}{ccccccccc}
n'_{.1} & = & n'_{1.} \frac{n_{11}}{n_{1.}} & + & n'_{2.} \frac{n_{21}}{n_{2.}} & + & \dots & + & n'_{i.} \frac{n_{i1}}{n_{i.}} \\
n'_{.2} & = & n'_{1.} \frac{n_{12}}{n_{1.}} & + & n'_{2.} \frac{n_{22}}{n_{2.}} & + & \dots & + & n'_{i.} \frac{n_{i2}}{n_{i.}} \\
\dots & = & \dots & + & \dots & + & \dots & + & \dots \\
n'_{.i} & = & n'_{1.} \frac{n_{1i}}{n_{1.}} & + & n'_{2.} \frac{n_{2i}}{n_{2.}} & + & \dots & + & n'_{i.} \frac{n_{ii}}{n_{i.}}
\end{array}\right\} \quad (7)$$

$$\begin{pmatrix} n'_{1.} \\ n'_{2.} \\ \dots \\ n'_{i.} \end{pmatrix} = \begin{pmatrix} \frac{n_{11}}{n_{1.}} & \frac{n_{21}}{n_{2.}} & \dots & \frac{n_{i1}}{n_{i.}} \\ \frac{n_{12}}{n_{1.}} & \frac{n_{22}}{n_{2.}} & \dots & \frac{n_{i2}}{n_{i.}} \\ \dots & \dots & \dots & \dots \\ \frac{n_{1i}}{n_{1.}} & \frac{n_{2i}}{n_{2.}} & \dots & \frac{n_{ii}}{n_{i.}} \end{pmatrix}^{-1} \begin{pmatrix} n'_{.1} \\ n'_{.2} \\ \dots \\ n'_{.i} \end{pmatrix} \quad (8)$$

Equation 8. The PERLE bias correction method.

TABLE IV. NOTATION USED IN EQUATION (8)

		True Class				Output Count
		C_1	C_2	...	C_i	
Output Class	C_1	n_{11}	n_{21}	...	n_{i1}	$n_{.1}$
	C_2	n_{12}	n_{22}	...	n_{i2}	$n_{.2}$
	...	...	...	...	...	...
	C_i	n_{1i}	n_{2i}	...	n_{ii}	$n_{.i}$
True Count		$n_{1.}$	$n_{2.}$	...	$n_{i.}$	

TABLE V. NUMERICAL EXAMPLE OF BIASES CORRECTED WITH (8)

		True Class			*Counts*		
Dataset	**Class**	**C1**	**C2**	**C3**	**Output**	**True**	**Corrected**
Ground-Truth	*C1*	80	10	0	90	100	-
	C2	15	85	10	110	100	
	C3	5	5	90	100	100	
Input Set1	*C1*	160	10	0	170	200	200
	C2	30	85	30	145	100	100
	C3	10	5	270	285	300	300
Input Set2	*C1*	80	10	1	91	100	100
	C2	15	85	19	119	100	99
	C3	5	5	180	190	200	206

Assessing uncertainty propagation - To account for *Object Detection* and *Tracking Errors* propagated to *Species Recognition Errors*, groundtruth evaluations can evaluate *Species Recognition Errors* by including a class for *unknown species* to represent the FP objects. This approach is applicable to other use cases applying species recognition prior to tracking algorithms, as groundtruth evaluations are performed for the whole pipeline of object detection, tracking and species recognition as a black box, whatever the order of the algorithms' sequence. Similarly, *Behavior Recognition Errors* can include a class *unknown behavior*. Classes for unknown species and behavior can be used to apply the bias correction method in (8), while accounting for uncertainty propagation.

With this approach, 2 additional aspects need to be handled: i) *Image Quality* and ii) *Sampling Coverage* and *Fragmentary Processing*. The latter can be handled by using equations (2-6), after correcting the counts of individuals with equation (8). Our strategy for handling *Image Quality* consists of labelling each video sample with an image quality class. Groundtruth evaluations can be repeated for image quality class, at the cost of requiring extensive groundtruth sets (i.e., sets of items for each combination of species, behavior and image quality). If magnitudes of misclassifications are significantly different, the bias correction method in (8) needs to be applied for each class of image quality. A numerical example is shown in Table VI.

TABLE VI. NUMERICAL EXAMPLE OF BIASES CORRECTION HANDLING IMAGE QUALITY

		True Class			*Counts*		
Dataset	**Class**	**C1**	**C2**	**C3**	**Output**	**True**	**Corrected**
Without Accounting for Image Quality							
Groundtruth All Images	*C1*	145	25	5	175	200	-
	C2	40	165	30	235	200	
	C3	15	10	165	190	200	
Input Set All Images	*C1*	355	40	5	400	500	487
	C2	105	245	50	400	300	290
	C3	40	15	345	400	400	423
Accounting for Image Quality							
Groundtruth Blurred Img.	*C1*	65	15	5	85	100	-
	C2	25	80	20	125	100	
	C3	10	5	75	90	100	
Groundtruth Normal Img.	*C1*	80	10	0	90	100	-
	C2	15	85	10	110	100	
	C3	5	5	90	100	100	
Input Set Blurred Img.	*C1*	195	30	5	230	300	300
	C2	75	160	20	255	200	200
	C3	30	10	75	115	100	100
Input Set Normal Img.	*C1*	160	10	0	170	200	200
	C2	30	85	30	145	100	100
	C3	10	5	270	285	300	300
Sum of Both Input Sets	*C1*	355	40	5	400	500	500
	C2	105	245	50	400	300	300
	C3	40	15	345	400	400	400

Our proposed framework to estimate *Biases Emerging from Noise*, and assess the *Uncertainty in Specific Output*, is summarized in Fig. 6. It relies on the assumption that errors measured in groundtruth evaluations are representative of errors occurring in subsets of computer vision outputs. However, stochastic variations of error magnitudes impact the correction of biases, as shown in Table V (Input Set 2). Further work is needed to control this assumption (e.g., measuring the variability of error magnitudes over random splits of groundtruth sets, and deriving confidence intervals for the corrected counts of individuals). Future work also needs to develop methods to i) estimate the number of groundtruth items needed for the bias correction method to be reliable; and ii) investigate methods to identify biases due to *Image Quality* while avoiding to collect extensive groundtruth sets representing all combinations of image quality, species and behavior.

VI. CONCLUSION

We have presented foundations for assessing the multifactorial uncertainty in computer vision systems used for monitoring populations in their natural environment. Our approach provides an original, comprehensive and interoperable set of measures allowing the application of standard statistical techniques for performing uncertainty-aware analyses of computer vision data. Future work needs to investigate the validity of our approach with empirical evaluations. However, this work provides an overview of uncertainty factors and assessment methods, and highlights issues that were unaddressed and prevented scientifically valid analysis of computer vision results.

ACKNOWLEDGMENTS

We are grateful for the invaluable contribution of Arjen P. de Vries in our effort to formalize the PERLE method.

REFERENCES

[1] K. J. Gaston and M. A. O'Neill, "Automated species identification: why not?" *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 359, no. 1444, pp. 655–667, 2004.

[2] M. A. Acevedo, C. J. Corrada-Bravo, H. Corrada-Bravo, L. J. Villanueva-Rivera, and T. M. Aide, "Automated classification of bird and amphibian calls using machine learning: A comparison of methods," *Ecological Informatics*, vol. 4, no. 4, pp. 206–214, 2009.

[3] R. Kays, B. Kranstauber, P. Jansen, C. Carbone, M. Rowcliffe, T. Fountain, and S. Tilak, "Camera traps as sensor networks for monitoring animal communities," in *IEEE 34th Conference on Local Computer Networks (LCN)*, 2009, pp. 811–818.

[4] D. A. Lytle, G. Martínez-Muñoz, W. Zhang, N. Larios, L. Shapiro, R. Paasch, A. Moldenke, E. N. Mortensen, S. Todorovic, and T. G. Dietterich, "Automated processing and identification of benthic invertebrate samples," *Journal of the North American Benthological Society*, vol. 29, no. 3, pp. 867–874, 2010.

[5] C. Spampinato, V. Mezaris, J. van Ossenbruggen, and M. Cristani, "Workshops on Multimedia Analysis for Ecological Data," in *Proceedings of the 20th, 21st and 22nd ACM international conferences on Multimedia*, http://dl.acm.org/citation.cfm?doid=2390832 (2012), http://dl.acm.org/citation.cfm?doid=2509896 (2013), http://dl.acm.org/citation.cfm?doid=2661821 (2014).

[6] E. Beauxis-Aussalet, E. Arslanova, and L. Hardman, "Supporting non-experts' awareness of uncertainty: Negative effects of simple visualizations in multiple views." in *Proceedings of the 33rd European Conference on Cognitive Ergonomics (ECCE)*. ACM, 2015.

[7] B. J. Boom, P. X. Huang, C. Beyan, C. Spampinato, S. Palazzo, J. He, E. Beauxis-Aussalet, S.-I. Lin, H.-M. Chou, G. Nadarajan *et al.*, "Long-term underwater camera surveillance for monitoring and analysis of fish populations," in *Workshop on Visual observation and analysis of Animal and Insect Behaviour (VAIB), held at the 21st International Conference on Pattern Recognition (ICPR)*, 2012.

[8] W. E. Walker, P. Harremoës, J. Rotmans, J. P. van der Sluijs, M. B. van Asselt, P. Janssen, and M. P. Krayer von Krauss, "Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support," *Integrated assessment*, vol. 4, no. 1, pp. 5–17, 2003.

[9] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler, "Visualizing geospatial information uncertainty: What we know and what we need to know," *Cartography and Geographic Information Science*, vol. 32, no. 3, pp. 139–160, 2005.

[10] A. T. Pang, C. M. Wittenbrink, and S. K. Lodha, "Approaches to uncertainty visualization," *The Visual Computer*, vol. 13, no. 8, pp. 370–390, 1997.

[11] C. D. Correa, Y.-H. Chan, and K.-L. Ma, "A framework for uncertainty-aware visual analytics," in *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*. IEEE, 2009, pp. 51–58.

[12] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

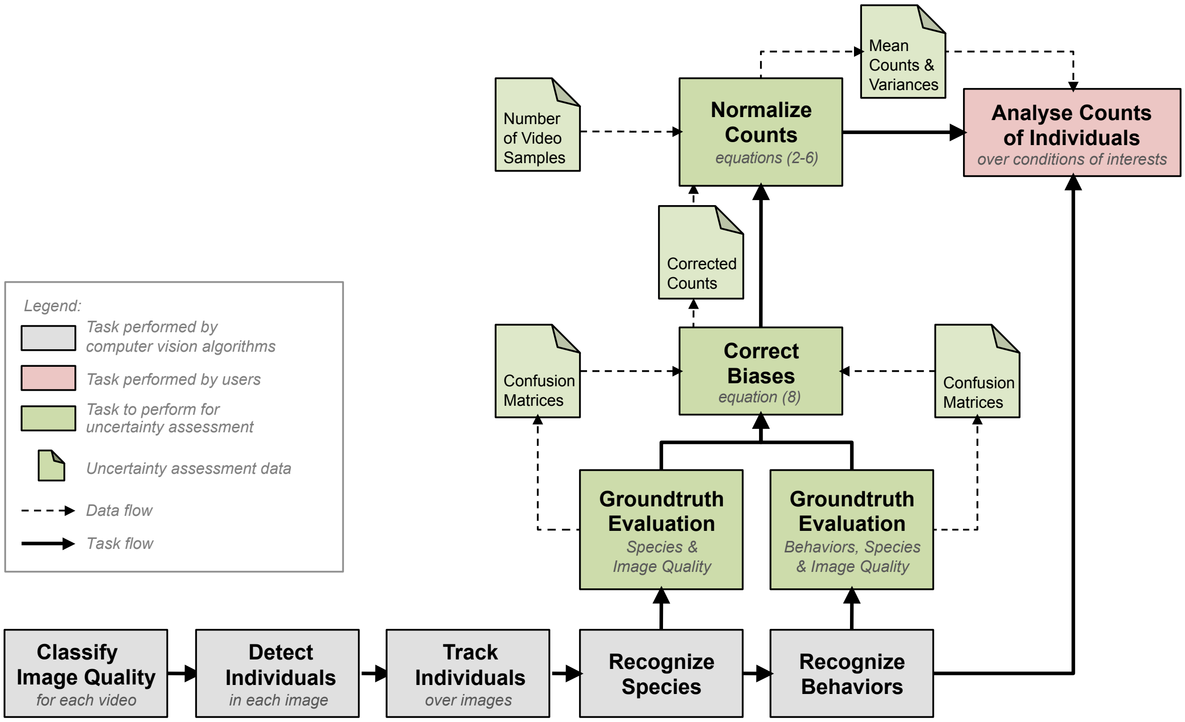

Fig. 6. Uncertainty assessment framework (complementing Fig. 3).

[13] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[14] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," *Lecture Notes in Computer Science*, vol. 1973, pp. 420–434, 2001.

[15] B. J. Boom, E. Beauxis-Aussalet, L. Hardman, and R. B. Fisher, "Uncertainty-aware estimation of population abundance using machine learning (in press)," *Multimedia System Journal*, 2015.

[16] C. Spampinato, S. Palazzo, D. Giordano, I. Kavasidis, F.-P. Lin, and Y.-T. Lin, "Covariance based fish tracking in real-life underwater environment," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2012, pp. 409–414.

[17] M. Crosetto, J. A. M. Ruiz, and B. Crippa, "Uncertainty propagation in models driven by remotely sensed data," *Remote Sensing of Environment*, vol. 76, no. 3, pp. 373–385, 2001.

[18] X. Zhu and X. Wu, "Class noise vs. attribute noise: A quantitative study," *Artificial Intelligence Review*, vol. 22, no. 3, pp. 177–210, 2004.

[19] R. Senge, J. J. Del Coz, and E. Hüllermeier, "On the problem of error propagation in classifier chains for multi-label classification," in *Data Analysis, Machine Learning and Knowledge Discovery*. Springer, 2014, pp. 163–170.

[20] N. J. Hetrick, K. M. Simms, M. P. Plumb, and J. P. Larson, *Feasibility of using video technology to estimate salmon escapement in the Ongivinuk River, a clear-water tributary of the Togiak River*. US Fish and Wildlife Service, King Salmon Fish and Wildlife Field Office, 2004.

[21] T. Yoshida, K. Akagi, T. Toda, M. Kushairi, A. Kee, and B. Othman, "Evaluation of fish behaviour and aggregation by underwater videography in an artificial reef in tioman island, malaysia," *Sains Malaysiana*, vol. 39, no. 3, pp. 395–403, 2010.

[22] D. L. Watson, E. S. Harvey, M. J. Anderson, and G. A. Kendrick, "A comparison of temperate reef fish assemblages recorded by three underwater stereo-video techniques," *Marine Biology*, vol. 148, no. 2, pp. 415–425, 2005.

[23] T. J. Willis and R. C. Babcock, "A baited underwater video system for the determination of relative density of carnivorous reef fish," *Marine and Freshwater Research*, vol. 51, no. 8, pp. 755–763, 2000.

[24] J. Irvine, B. Ward, P. Teti, and N. Cousens, "Evaluation of a method to count and measure live salmonids in the field with a video camera and computer," *North American Journal of Fisheries Management*, vol. 11, no. 1, pp. 20–26, 1991.

[25] I. Tulp, L. J. Bolle, and A. D. Rijnsdorp, "Signals from the shallows: in search of common patterns in long-term trends in dutch estuarine and coastal fish," *Journal of Sea Research*, vol. 60, no. 1, pp. 54–73, 2008.

[26] H. Visser, "Estimation and detection of flexible trends," *Atmospheric Environment*, vol. 38, no. 25, pp. 4135–4145, 2004.

[27] E. Beauxis-Aussalet, E. Arslanova, J. Van Ossenbruggen, and L. Hardman, "A case study of trust issues in scientific video collections," in *Proceedings of the 2nd ACM international workshop on Multimedia analysis for ecological data*. ACM, 2013, pp. 41–46.

[28] M. Edwards and D. R. Morse, "The potential for computer-aided identification in biodiversity research," *Trends in ecology & evolution*, vol. 10, no. 4, pp. 153–158, 1995.

[29] R. B. Fisher, B. J. Boom, P. X. Huang, C. Beyan, C. Spampinato, S. Palazzo, J. He, E. Beauxis-Aussalet, S.-I. Lin, H.-M. Chou, G. Nadarajan *et al.*, "Fish4Knowledge Project Deliverables," Tech. Rep., 2013. [Online]. Available: http://groups.inf.ed.ac.uk/f4k/deliverables.htm

[30] E. Beauxis-Aussalet and L. Hardman, "Visualization of confusion matrix for non-expert users," in *Poster at the IEEE Symposium on Information Visualization (IEEE VIS)*, 2014.

[31] J. He, J. van Ossenbruggen, and A. P. de Vries, "Do you need experts in the crowd?: a case study in image annotation for marine biology," in *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, 2013, pp. 57–60.

[32] E. Beauxis-Aussalet and L. Hardman, "Simplifying the visualization of confusion matrix," in *26th Benelux Conference on Artificial Intelligence (BNAIC)*, 2014.

[33] ——, "Multi-purpose exploration of uncertain data for the video monitoring of animal populations," in *Workshop on Visualization in Environmental Sciences (EnvirVis). Eurographics Association.*, 2015.

[34] E. Beauxis-Aussalet, J. Van Doorn, and L. Hardman, "Bridging the gap between machine learning experts and end-users with interactive uncertainty visualization," in *Poster at EuroVis*, 2015.

[35] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[36] W. G. Cochran, *Sampling techniques*. John Wiley & Sons, 2007.