Understanding Uncertainty Issues in the Exploration of Fish Counts

Emma Beauxis-Aussalet and Lynda Hardman

Abstract Several data analysis steps are required for understanding computer vision results and drawing conclusions about the actual trends in the fish populations. Particular attention must be drawn to the potential errors that can impact the scientific validity of end-results. This chapter discusses the means for ecologists to investigate the uncertainty in computer vision results. We address a set of uncertainty factors identified by interviewing both ecology and computer vision experts, as discussed in Chapter **??**. We investigate state-of-the-art methods to specify these uncertainty factors. We identify issues with conveying the results of ground-truth evaluation methods to end-users who are not familiar with computer vision technology, and we present a novel visualization design addressing these issues. Finally, we discuss the uncertainty factors for which evaluation methods require further research.

1 Introduction

As scientists, ecologists have requirements of transparency regarding the data collection process and its potential errors and biases. There are several uncertainty factors that potentially impact computer vision end-results, as discussed in Chapter **??**. Each uncertainty factor has specific effects on end-results, hence requiring specific evaluation methods. We interviewed both marine ecology experts and computer vision experts to gain insights on the effects of uncertainty factors, and on the methods for measuring them. In this chapter, we detail the potential effects of each uncertainty factor, the goals of their evaluation, the state-of-the-art evaluation methods, and the uncertainty visualizations developed within the project. Sections 2-3 investigate uncertainty related to computer vision algorithms, while sections 4-5 investigate uncertainty related to the in-situ deployment of the Fish4Knowledge

Emma Beauxis-Aussalet e-mail: emmanuelle.beauxis-aussalet@cwi.nl
Lynda Hardman e-mail: lynda.hardman@cwi.nl\atCWI, Sciencepark123, 1098XGAmsterdam, TheNetherlands

system. Section 6 investigates the impact of both computer vision algorithms and in-situ system deployment uncertainties on end-results. Finally, section 7 discusses the uncertainty issues that are not fully addressed by the state-of-the-art evaluation methods.

We show that the Fish4Knowledge project is supported by well-established methods for evaluating the uncertainty factors due to computer vision algorithms. The evaluation of the remaining uncertainty factors requires methods beyond the state-of-the-art. However, the Fish4Knowledge project developed simple evaluation methods for these factors. Directions for future work are suggested with the aim of enabling further scientific rigor in ecology research based on computer vision systems. An overview of the uncertainty factors and their evaluation methods is given by Figure 1 and Table 1. The latter refers to the user interface designed to communicate computer vision results and their uncertainty to end-users. The interface organises information in 5 tabs addressing specific uncertainty issues, and is further discussed in Appendix **??**.

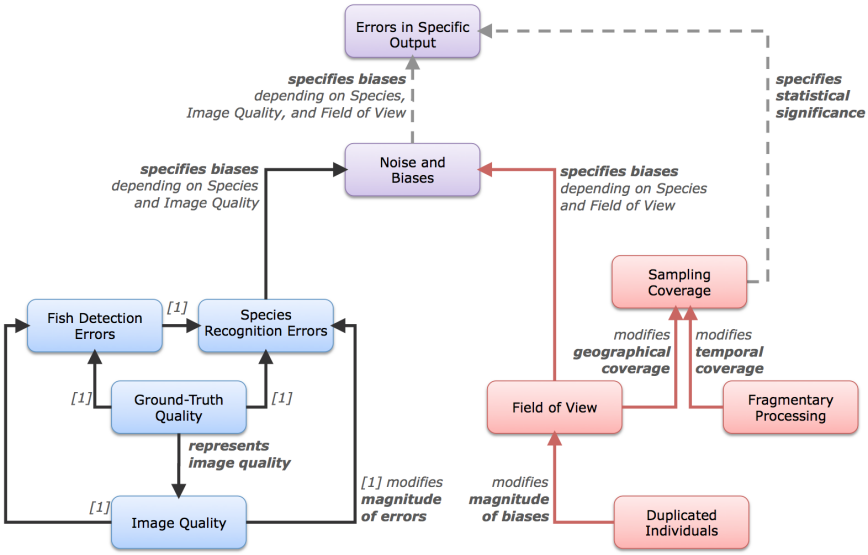

Fig. 1: Model of interactions between uncertainty factors. Factors in blue boxes are introduced by computer vision algorithms, while factors in red boxes are introduced when deploying the system. Factors in purple boxes are impacted by both phases of system implementation.

Table 1: Uncertainty factors and user interface tabs addressing them.

Factor	*User Interface Tab & Metrics*	*Figure*
Uncertainty due to computer vision algorithms (Sections 2-3)		
Ground-Truth Quality	**Video Analysis tab:** Number of ground-truth items in test sets (which are proportional to numbers of items in training sets).	Fig. 3-7
Image Quality	**Video** and **Visualization tabs:** Number of videos from each image quality (bottom widget called *Video Quality*) to correlate with *Fish Detection Errors*.	Fig. 8-9
Fish Detection Errors	**Video Analysis tab:** Number and proportion of errors per *Image Quality*.	Fig. 3-6
Species Recognition Errors	**Video Analysis tab:** Number and proportion of errors per species.	Fig. 4-7
Uncertainty due to in-situ system deployment (Sections 4-5)		
Field of View	**Video tab:** Video browsing supports elementary control of Fields of View over time and locations.	Fig. 8
Duplicated Individuals	**Video tab:** Video browsing supports elementary control of repeated occurrences of fish in groups or coral heads (e.g., over-estimation of *schooling* and *sedentary* species discussed in chapter **??**).	Fig. 8
Sampling Coverage	**Video** and **Visualization tabs:** Number of 10-minute video samples over time and locations.	Fig. 8-9
Fragmentary Processing	**Visualization tab:** Number of processed and unprocessed 10-minute video samples. Mean number of fish per video sample. Additional video processing can be requested through the user interface (**Workflow sub-tab**, Appendix **??**, Fig. **??** p.**??**).	Fig. 10
Uncertainty due to both computer vision algorithms and deployment conditions (Section 6)		
Noise and Biases	**Video** and **Video Analysis tabs:** Video browser and visualization of computer vision errors, to identify potential biases due to *Field of View, Duplicated Individuals, Image Quality, Fish Detection* and *Species Recognition Errors*.	Fig. 11
Uncertainty in Specific Output	**Visualization tab:** Measures of dataset characteristics, to correlate with *Noise and Biases* estimates (number of videos over time, location, *Image Quality* and *Field of View*), and *Certainty scores* indicating the similarity of fish with their species model. **Report tab:** Uncertainty can be described by gathering and commenting visualizations.	Fig. 10-11

2 Evaluating Uncertainty Due to Computer Vision Algorithms

Computer vision algorithms can introduce errors in end-results by misidentifying fish and non-fish objects, or by misidentifying fish species. To convey this uncertainty to end-users, we consider the two stages of information extraction as two distinct algorithms: *Fish Detection* for identifying fish from other objects (Chapters **??**-**??**), and *Species Recognition* for identifying the fish species (Chapter **??**). Besides algorithms themselves, two factors can impact the quality of the output. Algorithms use ground-truth sets of fish examples to learn how to identify fish from each species. The *Ground-Truth Quality* directly impacts the quality of end-results. Further, the *Image Quality* of video recordings can induce errors, e.g., low image

quality yields fish appearances that are more difficult to recognize. The interactions between these uncertainty factors are shown in Figure 1 (blue boxes). In this section, we present these uncertainty factors and their evaluation methods.

Fish Detection and Species Recognition Errors - Computer vision algorithms identify the fish appearing in video footage by classifying them into predefined categories. The *Fish Detection* algorithm has two categories, fish or non-fish objects, and *Species Recognition* has one category for each fish species. For both algorithms, objects are assigned to a single category. The fish from the *Fish Detection* results are classified further by *Species Recognition*.

Each fish category is defined by a model constructed from ground-truth sets. Objects are compared to the models, and if similar enough, are classified in the related categories. Similarity between objects and models is usually represented with a *score*. *Score* thresholds are used for selecting the objects to classify, and are usually set by computer vision experts. Errors occur when objects are not classified into their true category, or when they are not classified at all (i.e., not detected in the videos).

Fish Detection output is impacted by two types of errors. Errors of *Type I*, also called *False Positives* (FP), are non-fish objects classified as fish and contained in the output. Errors of *Type II*, also called *False Negatives* (FN), are undetected fish not contained in the output. These errors are usually measured using ground-truth sets distinct from those used to learn the fish models. Manual fish detections are compared to those of the algorithm, and the numbers of errors are encoded in a table called a *confusion matrix*. Table 2 illustrates a typical confusion matrix for *Fish Detection Errors*.

Species Recognition Errors are fish that have been assigned to the wrong species. They are also measured using dedicated ground-truth sets, and encoded in a confusion matrix. Table 3 shows an example of a confusion matrix for *Species Recognition Errors*. Type I and II errors also apply to *Species Recognition*. Considering a set of fish assigned to one species, e.g., *Species A*, Errors of *Type I* (*False Positives*) are fish from another species erroneously classified as *Species A*. Errors of *Type II* (*False Negatives*) are fish not classified as *Species A* but actually belonging to it.

Confusion matrices for *Species Recognition* are more complex to analyze than those of *Fish Detection*. An important concept for understanding them is that False Positives erroneously assigned to one species are False Negatives for their true species. For instance, if *Species A* misses 17 False Negatives erroneously attributed to *Species B*, then *Species B* gains 17 False Positives from *Species A*. Hence counting all the errors for one species requires to sum the False Negatives assigned to all other species (i.e., column-wise sum in Table 3), as well as summing the False Positives added by all other species (i.e., row-wise sum in Table 3). This examples is illustrated in Table 3, e.g., the cell with both red and grey squares indicates: 17 False Negatives (FN) for species A; 17 False Positives (FP) for species B. These 17 errors are counted both in the cell with red background (i.e., summing the cells with red squares) and in the cell with grey background (i.e., summing the cells with grey squares).

		Classification from Ground-Truth	
		Fish	Non-Fish
Classification from Fish Detection Software	Fish	**85 (True Positives TP)**	7 (False Positives FP)
	Non-Fish	**15 (False Negatives FN)**	93 (True Negatives TN)

Table 2: Example of a confusion matrix for *Fish Detection Errors* (with synthetic data). The color coding is used in our visualization design to facilitate the identification of type I and II errors.

		Classification from Ground-Truth					Basic Metrics			
		A	B	C	D	E	TP	FN	FP	*TN*
Classification from Species Recognition Software	A	**85**	1	4	3	12	**85**	**25**	20	384
	B	17	**78**	1	7	2	**78**	17	**27**	392
	C	1	2	**90**	6	6	**90**	22	15	387
	D	5	7	2	**77**	1	**77**	18	15	404
	E	2	7	15	2	**81**	**81**	21	26	386

Table 3: Example of a confusion matrix with synthetic data for *Species Recognition Errors* (left) and basic metrics for type I and II errors (i.e., FP and FN, respectively).

In the computer vision domain, classification errors are usually synthesized further using advanced metrics derived from the basic measure of TP, FP, FN and TN. Advanced metrics are rates of correct and incorrect object detection over total numbers of objects belonging to the category (TP and FN) or not (FP and TN). Table 4 shows the metrics and formulas commonly used in most of the state-of-the-art evaluations of computer vision errors. Advanced metrics are usually plotted by pairs in Precision/Recall or ROC curves (Receiver Operating Characteristics). Measurements are usually repeated for several parameter thresholds, e.g., a *score* representing the similarity between fish images and species models (i.e., fish below thresholds are discarded from the species). Figure 2 shows examples of such visualization of pairs of advanced metrics.

Precision	Recall or TP Rate	FP Rate	Accuracy	F1 Measure
$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$\frac{FP}{FP+TN}$	$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{2TP}{2TP+FP+FN}$

Table 4: Advanced metrics commonly used in computer vision.

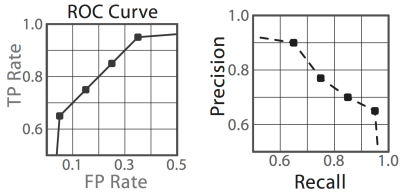

Fig. 2: Example of an ROC (left) and Precision/Recall curve (right). Error rates are given for different parameter settings, e.g., the points on the curves show 4 *score* thresholds discarding objects that are not similar enough to the fish model.

Image Quality - Varying image quality can be a source of bias. For instance, end-results drawn from one type of image quality can systematically contain different numbers of errors than for another image quality. This biases the comparison of end-results drawn from different types of image quality. Ground-truth evaluations of *Fish Detection* and *Species Recognition Errors* can be used to evaluate this type of bias.

Hence we need to provide ecologists with evaluations of *Fish Detection* and *Species Recognition Errors* detailed for each type of image quality. However, this requires an extensive ground-truth containing sufficient numbers of annotations for all combinations of species and image quality. The considerable cost of such ground-truth collection is likely to be unaffordable, as it was the case for the Fish4Knowledge project. *Species Recognition Errors* could not be fully evaluated for each image quality. Hence we focused on evaluating *Fish Detection Errors* for each type of *Image Quality*.

Image quality is automatically detected prior to *Fish Detection*, and specific parameter tuning is applied for adapting the computer vision algorithm to the characteristics of image quality. To investigate uncertainty due to image quality, ground-truth evaluation of *Fish Detection Errors* were performed for each *Image Quality*. When analysing fish counts from a set of video samples, users can relate the numbers of videos from each image quality with the errors measured for each image quality.

Ground-Truth Quality - Ground-truth sets contain examples of fish that were manually annotated by ecology experts, or by non-experts recruited from crowdsourcing (Chapter **??**). Computer vision algorithms learn to recognize fish and their species by constructing fish models on the basis of these examples. Hence ground-truth is essential to ensure the quality of information extraction. Issues arise with ground-truth sets that are not representative of the possible fish appearances, and with scarcity of fish examples, e.g., for rare species. To be representative of the fish populations, ground-truth sets need to contain examples of the typical fish appearances. For instance, if a species color can vary between grey and black, the ground-truth must contain examples of both grey and black appearances. Similarly, if cameras often record blurred and low-contrast images, then the ground-truth should con-

tain examples of fish for each image quality. This is usually ensured by selecting ground-truth images through a random sampling among all images collected from all cameras.

Ground-truth can contain outliers such as erroneous annotations, or rare fish appearances (e.g., odd fish poses). With scarce ground-truth, outliers can have a great impact on computer vision errors. For instance, if a small ground-truth set contain an image of seaweed, then the fish model can be distorted so as to be compatible with seaweed appearances. Hence a high number of non-fish objects can be included in end-results. Large ground-truth sets potentially lower the impact of outliers, as outliers' distortion of fish models is likely to be overridden by numerous counter examples.

Hence, to evaluate uncertainty due to ground-truth quality, we need to measure the representativity of ground-truth sets, their annotation errors, and the quantity of ground-truth items. Ground-truth quantity is the number of examples of each type of fish to recognize: examples of fish and non-fish objects for the *Fish Detection* algorithm, and examples of each species for the *Species Recognition* algorithm. Regarding annotation errors, several metrics exist: number of annotators for each image, level of expertise of annotators (e.g., professor, student, or inexperienced), and level of agreement amongst annotators if annotations are contradictory (e.g., Cohen's kappa). They are typically applied for evaluating ground-truth sets collected through crowd-sourcing (Chapter **??**). For ground-truth representativity, we need to take into account the *Image Quality* of the recordings. The number of ground-truth items for each image quality indicates potential scarcity for one type of image, which increases uncertainty in end-results drawn from such videos. A randomized selection from a large quantity of ground-truth items ensures *a priori* that ground-truth sets are representative of the fish appearances. This ground-truth collection method is recommended both for the ground-truth sets used for learning fish models, and the sets used for evaluating *Fish Detection* and *Species Recognition Errors*. However, future work is needed for formally assessing ground-truth representativity, and for assessing that sufficient numbers of items are collected.

During interviews with ecologists, we explained the ground-truth annotation process. Ecologists were interested in the numbers of ground-truth images, and in browsing them. Further metrics, such as numbers of annotators and their level of agreement, were not introduced at first to avoid overwhelming users. We focused on providing the numbers of ground-truth items correctly or incorrectly classified, for each species or image quality. Future work can investigate the benefits of providing further metrics to end-users, e.g., the level of agreement between annotators, to improve user confidence.

3 Visualizing Uncertainty Due to Computer Vision Algorithms

End-users who are not familiar with computer vision are likely to encounter difficulties in understanding ground-truth evaluations and their technical concepts [**?**]. Some metrics may be misunderstood or may not fully address the uncertainty factors. This section summarizes these issues, and presents a visualization design adapted for end-users who are not necessarily experts in computer vision.

3.1 Usability Issues with Computer Vision Evaluations

Confusion matrices need to be read both column- and row-wise, which is tedious and error prone. For instance, considering the cell with both red and grey squares in Table 3, if read row-wise it indicates False Positives added to *Species B*. If read column-wise, it indicates False Negatives lost by *Species A*. Memorizing all cell values, and their semantics, is an important cognitive load. Users may forget cell values, or may read only columns or rows.

To limit cognitive efforts, confusion matrices can be synthesized by cumulating errors for each species (i.e., basic metrics in Table 3). However, it is no longer possible to distinguish which species are likely to be confused with another. For instance, the cells with red or grey background in Table 3 do not indicate the original true species of the misrecognized fish. Users need this information to identify correlations between species populations that are induced by *Species Recognition Errors*, and hence, that are not representative of the actual trends in fish populations. For instance, an important increase of one species implies an increase of its False Negatives. A proportion of its fish are attributed to other species, and this can induce deceiving increases of other species, especially for species of much inferior abundance. Hence, users need to inspect errors between pairs of species, rather than the synthesis of False Positives and False Negatives cumulated for all species.

Finally, advanced metrics are more complex and convey specific types of errors, and thus non-expert users may misinterpret them. For instance, with *Species Recognition*, True Negatives are fish correctly discarded from a species. They are cumulated over all other species, and are usually of a much higher magnitude than True Positives, False Positives and False Negatives, as shown in Table 3. High numbers of True Negatives yield low *False Positive Rate* and high *Accuracy* (see formulas in Table 4). Hence this may conceal important numbers of False Positives or False Negatives. The visualizations commonly used by computer vision experts use pairs of advanced metrics (e.g., Figure 2). Considering the above-mentioned issues, such visualizations are likely to be overwhelming and misleading for end-users that are not familiar with computer vision. Moreover, advanced metrics no longer indicate the number of items in the ground-truth, and thus possible ground-truth scarcity. Confusion matrices originally provide this information, i.e., the numbers of test items which are usually proportional to the numbers of training items. Hence we investigated the ways to communicate numbers of test items correctly or incorrectly

classified, rather than the ways to communicate complex and potentially misleading error rates.

3.2 Preliminary User Study

Ecologists need to understand computer vision errors, but ground-truth evaluations techniques are complex and may overwhelm them. Hence we investigated which level of detail needs to be disclosed to end-users [**?**]. We exposed 7 marine ecologists to explanations progressively disclosing the concepts of ground-truth evaluations. Explanations were given in 3 steps. Each step consisted of i) a visualization of fish counts, as produced by our computer vision system; ii) a visualization of a ground-truth evaluation of our system, introducing new technical concepts; and iii) a questionnaire evaluating the impact of the new details introduced. The first step introduced the concepts of ground-truth sets used for training and evaluating the video analysis software. The uncertainty visualization simply compared manual and automatic fish counts. The second step introduced the concepts of True Positive (TP), False Negative (FN) and False Positive (FP). The uncertainty visualization showed manual and automatic fish counts, with details about the amount of TP, FN and FP. As a first step, True Negative were omitted to avoid overwhelming users. The third step introduced the concepts of fish model and *score thresholds* of classifiers. The *scores* measure how fish images look like the *fish models*, as discussed in section 2. The visualization presented sets of fish counts produced by using different *score thresholds*, and ground-truth evaluation of TP, FN, and FP given for each *score threshold*.

	Trust			**Acceptance**			**Understanding**			**Info. Needs**		
Step:	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
User 1	+	+	-	+	+	+	-	-	-	- -	- -	- -
User 2	-	-	+	+	+	+	+	-	+	- -	-	-
User 3	++	++	++	++	++	++	+	- -	- -	-	-	- -
User 4	-	-	-	-	-	-	++	++	++	-	-	-
User 5	- -	- -	-	+	+	++	++	++	++	- -	- -	- -
User 6	- -	- -	-	- -	-	-	+	+	-	- -	- -	- -
User 7	+	-	+	+	+	++	- -	- -	-	- -	+	-

Table 5: Qualitative analysis of the experiment introducing technical concepts of ground-truth based evaluations. *The quality of user trust, acceptance, understanding, and satisfaction of information needs is either Very High (**++**), High (**+**), Low (**-**), or Very Low (**- -**). Green cells indicate a positive effect of the explanation steps, orange cells indicate a negative effect, uncolored cells indicate no significant effect.*

At each step, a questionnaire measured i) *user trust* in the computer vision software's ability to count fish; ii) *user acceptance* of the software for scientific re-

search; iii) *user understanding* of the technical concepts; and iv) the satisfaction of *user information needs* for uncertainty evaluation. The questionnaires were independently analyzed by two experts in Human-Computer Interfaces. A 4-grade scale was used to (*Very Low - -*, *Low -*, *High +*, *Very High ++*) to qualify user trust, acceptance understanding and information needs. Table **??** summarizes the results of this experiment. It shows that the technical concepts were generally difficult to understand. Extensive time and additional explanations were required for ecologists to familiarize with them. Further, users information needs were not fully satisfied. For instance, users required to watch videos themselves and to inspect other uncertainty factors. Besides these issues, user acceptance remained globally unchanged over explanation steps. User acceptance is relatively high since computer vision can greatly reduce material costs and human efforts. The third step, introducing *score thresholds*, had a slightly positive impact on user trust and acceptance, i.e., in respectively 4 and 2 cases out of the 6 cases that could be improved (User 3 already had maximum score).

3.3 Visualization Design for Non-Expert Users

We designed visualizations intended to limit cognitive load and misunderstandings, while addressing the 4 uncertainty factors related to computer vision algorithms. Our first design choice is to omit the True Negatives. They are not contained, and should not be contained, in end-results as they are not informative from a user viewpoint. Further, [**?**] shows that understanding the concepts of True Positive, False Negatives and False Positives is already likely to overwhelm users. Finally, as the magnitude of True Negatives can largely exceed that of errors (False Positives and False Negatives), True Negatives may conceal uncertainty (see section 3.1).

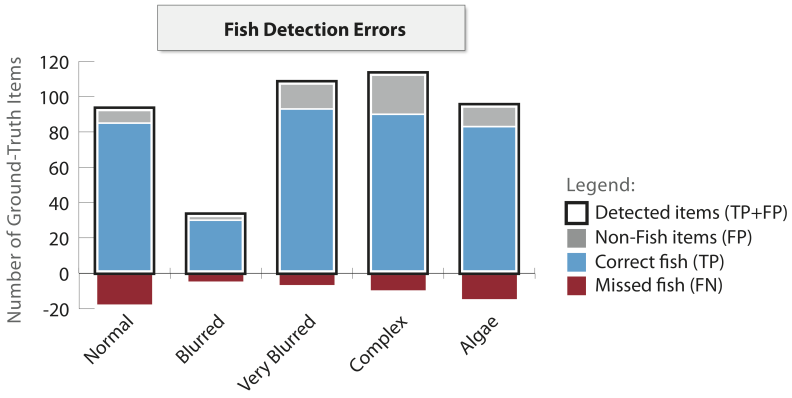

Fig. 3: Example of our novel visualization design detailing *Fish Detection Errors* for each type of *Image Quality*.

To avoid further cognitive load and misunderstandings, we avoid advanced metrics (Table 4). Our visualizations primarily show the numbers of ground-truth items yielding True Positives, False Negatives or False Positives in computer vision results. Figure 3 gives an example of such a display for *Fish Detection Errors*. It shows the numbers of ground-truth items, an important aspect of *Ground-Truth Quality* which is abstracted in traditional ROC or Precision/Recall curves.

The layout of our visualization intends to intuitively convey the concepts of *correct fish* (i.e., True Positives), *missed fish* (i.e., False Negatives) and *added fish* (i.e., False Positives). Stacked charts show the fish contained in end-results above a horizontal line, with *correct fish* below and *added fish* on top. *Missed fish* are displayed below the horizontal line. Colors reinforce the perception of errors. *Correct fish* are shown in blue, a positive or neutral color. *Added fish* are shown in light grey, to express an elusive presence contrasting with the saturated blue of *correct fish*. *Missed fish* are shown in red, a negative color expressing a warning. It aims at creating an intuitive perception that *missed fish* below the line are not included in end-results, and that *added fish* create over-estimations.

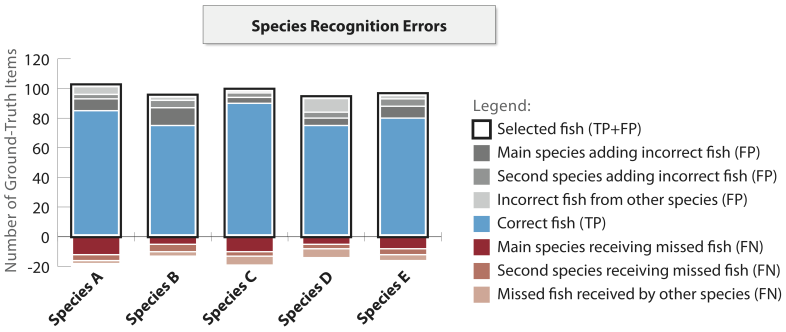

Fig. 4: Example of our novel visualization design for *Species Recognition Errors*.

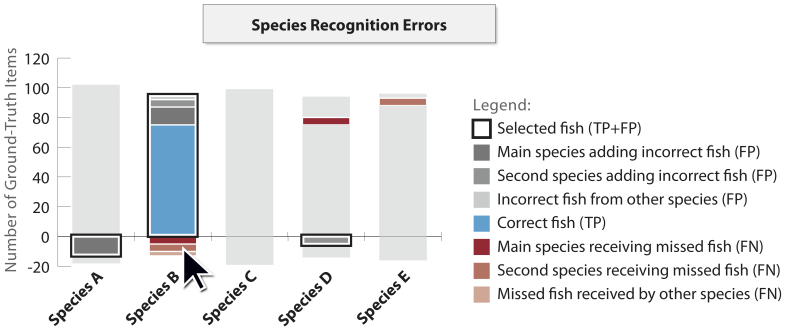

Fig. 5: Selecting a species of interest highlights the errors for that species. It shows from which species its False Positives (FP) came from (grey stacked bars) and to which species its False Negatives (FN) are attributed (red stacked bars).

For visualizing *Species Recognition Errors*, the same design principles are applicable. True Positives, False Positives and False Negatives can be displayed for each species. However, ecologists need to investigate which species are often confused with another, so as to identify potential biases with look-alike species. Hence, we need to detail which species adds False Positives, or receives False Negatives, and what is the magnitude of errors. Multiple confusions between pairs of species can occur, especially since errors are directional: e.g., fish from Species A misclassified as Species B ($FN_{a\rightarrow b}$), and inversely, fish from Species B misclassified as Species A ($FN_{b\rightarrow a}$). With N_s species $N_s(N_s-1)$ pairs of species need to be investigated. This complexity can clutter the visualization and overwhelm users. To address this issues, our visualization displays the most important inter-species confusions, and summarizes the remaining errors. For each species, we select the 2 other species yielding the most FP and FN, and display the related errors in distinct stacked block. The remaining errors from other species are displayed together in one block. Figure 4 gives an example of such display. Users can select a species to display errors only for that species, as shown in Figure 5.

The numbers of ground-truth items can greatly vary amongst classes of species and image quality, e.g., scarcity for some classes, or abundance of other classes. In these cases, ground-truth errors can be difficult to visualize. Hence users can switch between visualizing errors either as: i) numbers of ground-truth items; or ii) proportional measure of errors (Fig. 6-7). *Fish Detection Errors* are given proportionally to the total number of detected items ($TP+FP$), using equations (1,2). This choice of denominator intends to support the extrapolation of errors in subsets of end-results, for which only the total numbers of detected items are known. For instance, given a set of N_i fish detected in a set of videos with image quality Q_i, a user can extrapolate that it contains $N_i \frac{FP_i}{TP_i+FP_i}$ False Positives, and $N_i \frac{FN_i}{TP_i+FP_i}$ False Negatives (FP_i, FN_i and TP_i being measured from a ground-truth set representative of image quality Q_i).

$$Type\ I\ Error\ Rate\ Q_i = \frac{FP_i}{TP_i+FP_i} \tag{1}$$

$$Type\ II\ Error\ Rate\ Q_i = \frac{FN_i}{TP_i+FP_i} \tag{2}$$

Equations 1-2: *Type I* and *Type II Error Rates* Q_i are, respectively, the ratio of non-fish objects (FP_i) and undetected fish (FN_i) on the total numbers of detected items (TP_i+FP_i), measured in a ground-truth set of image quality Q_i. TP_i is the number of fish correctly detected for the ground-truth set. Equation (1) is equivalent to *Precision*, and equation (2) to *False Discovery Rate*.

For *Species Recognition Errors*, the False Negatives transferred from species *A* to species *B* ($FN_{a\rightarrow b}$), which are also the False Positives attributed to species B while truly belonging to species A ($FP_{b\leftarrow a}$), are given proportionally to the True Positive for species *A* using the equation (3). The choice of a denominator is different from that of error rates for *Fish Detection Errors* because in the case of *Species Recognition Errors*, False Positives and False Negatives are not independent between classes, i.e., between species. False Negatives for one species are False Positives for other species. Hence, in ground-truth evaluations, the number of False Positives

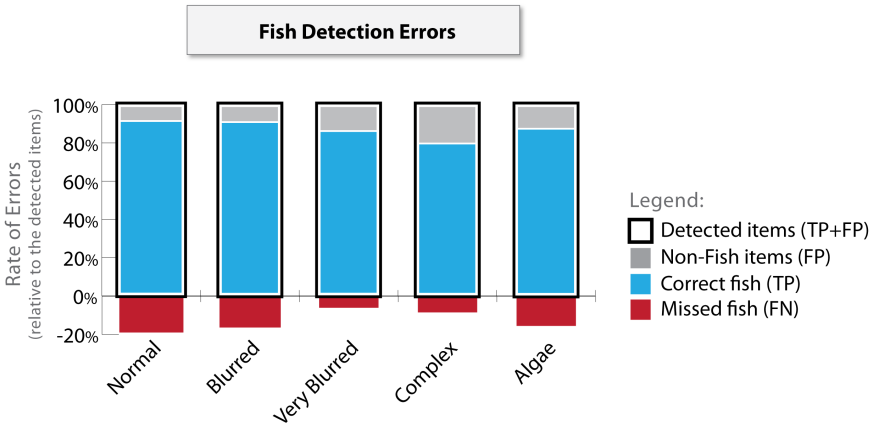

Fig. 6: Visualization design from Fig. 3 showing rates of errors from equations 1-2.

observed for one species arbitrarily varies depending on the number of ground-truth items for other species, independently from computer vision algorithms. Furthermore, each subset of end-results may have different proportions of each species, as population dynamics may be different over seasons or locations. For instance one species may be more abundant at specific periods of time than others. Hence, in end-results, the number of False Positives attributed to one species arbitrarily varies depending on the population sizes of other species, independently from computer vision algorithms. Therefore, False Positives are excluded from the denominator of error rate (3). The choice of a denominator also intends to support extrapolations of errors in subsets of end-results, as for the error rates (1-2). To do so, the denominator needs to represent the fish counts as observed in subsets of end-results, i.e., total numbers of fish detected for each species. Therefore, False Negatives are excluded from the denominators, as they are not contained in end-results' fish counts for each species. Using only True Positives as the denominator of error rates in equation (3) is a tradeoff between representing fish counts as observable in sets of end-results (i.e., $TP + FP$), and accounting for numbers of errors that are proportional to the population size of their true species (i.e., excluding FP which are not proportional to the size of their attributed species).

$$Pairwise\ Error\ Ratio\ S_a \rightarrow S_b = \frac{FN_{a \rightarrow b}}{TP_a} \tag{3}$$

Equation 3: *Pairwise Error Ratio* $S_a \rightarrow S_b$ is the ratio of fish belonging to species A (S_a) erroneously attributed to species B (S_b). $FN_{a \rightarrow b}$ is the number of ground-truth items attributed to S_b while truly belonging to S_a (e.g., the cell with both red and grey squares in Table 3). Note that $FN_{a \rightarrow b} = FP_{b \leftarrow a}$, i.e., the number of False Positives attributed to species B while truly belonging to species A. TP_a is the total number of TP for S_a. Note that $FN_{a \rightarrow b}$ is different from $FN_{b \rightarrow a}$ and *Pairwise Error Ratio* $S_a \rightarrow S_b$ is different from *Pairwise Error Ratio* $S_b \rightarrow S_a$.

To conclude, our visualization supports the evaluation of 4 uncertainty factors due to computer vision algorithms by showing a simple but complete representation of ground-truth evaluation results. *Fish Detection* and *Species Recognition Errors* are evaluated by visualizing absolute and relative numbers of errors in ground-truth sets (Figures 3-7). The uncertainty due to *Ground-Truth Quality* is evaluated by visualizing absolute numbers of ground-truth items, and the uncertainty due to *Image Quality* is evaluated by visualizing *Fish Detection Errors* for each type of image (Figure 3).

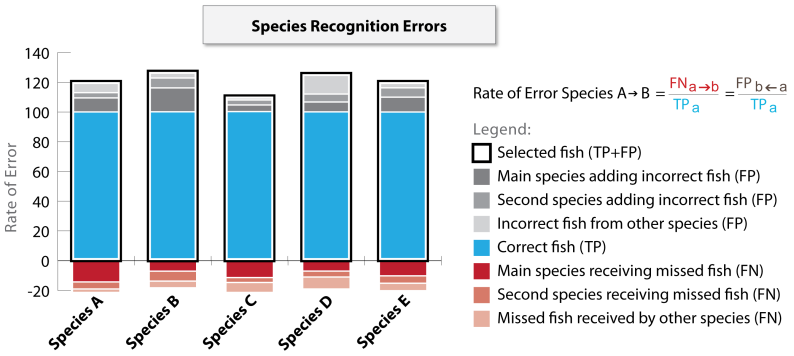

Fig. 7: Visualization design from Fig. 4 showing rates of error from equation 3.

4 Evaluating Uncertainty Due to In-Situ System Deployment

During our study of user requirements (Chapter **??**), we identified uncertainty factors that are not related to computer vision algorithms but to the deployment of the system. The deployment of cameras over marine ecosystems can greatly impact end-results, independently from potential computer vision errors. The cameras' field of view can increase or decrease the chances to observe specific species, hence creating biases. Some fields of view increase the chances of counting repeatedly the same individuals swimming back and forth in front of the camera, hence creating further biases. These types of biases typically concern *benthic* (i.e., living on the seabed), *sedentary* (i.e., living in coral heads), *schooling* (i.e., living in groups), and *herbivorous or carnivorous* species.

Furthermore, camera deployment over geographical locations may not provide a sufficient sampling of the ecosystem. Ecologists usually need redundant measurements to ensure the statistical validity of their observations [**?**]. Hence sufficient numbers of cameras need to be deployed. Additionally, the extent of the sampling coverage can be reduced if all collected videos are not processed due to technical issues. In this section, we discuss evaluation methods for these uncertainty factors related to the in-situ deployment of the Fish4Knowledge system.

Field of View and Duplicated Individuals - The different types of coral are populated by specific species feeding on its organisms, or hiding in its structure. Thus the placement of cameras needs to reflect the different habitats of interest. If some habitats are not observed, their species are under-represented, and end-results are biased. For instance, observations of *benthic, sedentary* and *carnivorous and herbivorous* species are biased if fields of view do not cover the specific habitats where these species are living.

Some species swimming behaviours (e.g., coming in and out of coral cavities) yield repeated occurrences of the same fish in the cameras' field of view. For instance, *schooling* and *sedentary* species (e.g., living in groups, or coral head cavities) are likely to yield *Duplicated Individuals* in end-results. Fields of view contribute to biases due to multiple re-identification of the same fish. For instance, close-ups on specific coral heads increase the chances of observing duplicated fish from *sedentary* species. Similarly, groups of *schooling* fish may not be consistently observed between close-ups and open sea views. Further, the depth of field of view modifies the sampling coverage of the area. For instance, compared to open sea views, close-up views cover a smaller area of the ecosystem.

The state-of-the-art does not offer well-established methods for handling these uncertainty factors. Future work needs to develop measures of rates of *Duplicated Individuals* depending on fish species and *Fields of View*. For example, a measure of such potential bias can indicate that *schooling* species S observed from field of view V are over-estimated by $FP_{s,v}\ Rate = \frac{FP_{s,v}}{TP_{s,v}+FP_{s,v}}$, similarly to error rate (1).

Finally, the Fish4Knowledge system relies on fixed cameras which fields of view are expected to remain the same over time. However, fields of view may vary over time. Small shifts can occur during maintenance and lens cleaning operations, and larger shifts can occur with environmental events such as typhoons. Hence, accidental changes of field of view need to be controlled and monitored over time.

Sampling Coverage and Fragmentary Processing - The Fish4Knowledge system stores continuous video footage into 10-minute excerpts. Ecologists need to take into account the number of 10-minute video samples from which computer vision results are drawn, since it influences the statistical representativity of the patterns observed in fish populations. For instance, fish counts observed from a few videos may not be representative of the actual populations of the ecosystem. Further, if different fish counts are drawn from video sets of different size, the more videos the more fish, hence comparison is biased. Therefore users need evaluations of sampling size (e.g., the numbers of videos over time periods and locations), and a comparable measure of fish abundance for end-results drawn from different sampling sizes.

The primary metric for sampling size is the number of video samples from which end-results are extracted. Additionally, the number of unprocessed videos, i.e., still in the workflow processing queue (Chapter **??**), indicates that further video processing could complement the end-results. The Fish4Knowledge system offers functionalities for manually requesting that specific videos of interest are processed with high priority [**?**].

To analyze sets of end-results extracted from varying numbers of video samples, averaging fish count per video as in (4) offers a comparable metric of fish abundance. Further, measuring variance over samples as in (5) complements the estimation of uncertainty. Such measure of variance over samples is often used as a basis for statistical analysis [**?**].

$$Mean\ Fish\ Count\ per\ Video = \frac{Number\ of\ Fish}{Number\ of\ Videos} \tag{4}$$

Equation 4: Measure of fish abundance for comparing fish counts drawn from varying numbers of videos.

$$Variance\ over\ Videos = \frac{1}{N_v}\sum_{i=1}^{N_v}(Mean\ Fish\ Count\ per\ Video - N_i)^2 \tag{5}$$

Equation 5: Measure of variance in fish abundance. N_v is the number of 10-min video sample, N_i is the number of fish in the i-th video sample.

However, the measures of mean and variance of fish counts per video in (4-5) must be used with care as they face three problems:

1) *Video duration* must be identical over samples. Video samples of longer duration are likely to contain more fish that samples of shorter durations, thus biasing the mean fish count per video as in (4). For video samples of unequal duration, fish abundance can be assessed by averaging fish counts over a time unit, e.g., $Mean\ Fish\ Count\ per\ Minute = \frac{1}{N_v}\sum_{i=1}^{N_v}\frac{Number\ of\ Fish\ in\ Video\ Sample\ i}{Duration\ of\ Video\ Sample\ i\ (in\ min)}$. We recommend the use of video samples of equal duration. Using videos of different durations would considerably complicate the measurement of uncertainty, particularly for analyzing the variance of fish abundance over different cameras while taking into account missing videos, as explained below.

$$Mean\ Abundance\ per\ 10\text{-}min = \sum_{j=1}^{N_c} Fish/Video\ at\ Camera\ C_j \tag{6}$$

Equation 7: Measure of fish abundance for analyzing fish counts drawn from several cameras, with varying numbers of video per camera. N_c is the number of cameras. $Fish/Video\ at\ Camera\ C_j = \frac{Number\ of\ Fish\ at\ Camera\ C_j}{Number\ of\ 10\text{-}min\ Videos\ at\ Camera\ C_j}$, i.e., the result of equation (4) for one camera.

2) *Fish abundance over different cameras*, and for the same time period, must be measured by summing the results of equation (4) for each camera separately, as in equation (7). It would be conceptually inaccurate to measure fish abundance as the result of equation (4) for all cameras globally, i.e., $\frac{Number\ of\ Fish\ for\ All\ Cameras}{Number\ of\ Videos\ for\ All\ Cameras}$. This is because the cameras observe the same time periods. For instance, if cameras 1 and 2 observe the same 10-minute time period, yielding 2 video samples with respectively N_1 and N_2 fish occurrences, then the overall fish abundance is $N_1 + N_2$ rather than $\frac{N_1+N_2}{2}$. To clarify what the metric represent, we recommend to use the label *Mean Abundance per 10-min* rather *Mean Fish Count per Video*. Note that

if the cameras' field of view observe the same overlapping areas, the overall fish abundance cannot be assessed as in (7).

3) *The variance of fish abundance over different cameras* (i.e., the variance of equation (7) results) is equivalent to the variance of a sum of random variables (the variables being the results of equation (4) for each camera). Such variance is measured by summing the covariances of equation (4) results over cameras, as in equation (8). Measuring such covariance assumes that video samples are available for all cameras, and for all the 10-min time periods. For instance, if *Mean Fish Counts per Video* at C_1 include the 10-min time period t_1 (e.g., 16:00 to 16:10 on Jan.1st 2011), then video samples must be available at all cameras for the same time period t_1. Then covariances can be measured using equation (9). If video samples are missing, i.e., if a 10-min time period is covered by at least 1 camera but not by all cameras, then it is not possible to measure covariances using equation (9). However, statistical methods can address this problem [**?**]. They consists of discarding incomplete sample subsets or using replacement values for missing samples (imputation). None of them can provide perfect results, and the choice of a method depend on each use case constraints.

$$Variance\ over\ Cameras = \sum_{j=1}^{N_c} \sum_{k=1}^{N_c} Cov(Fish/Video\ at\ C_j,\ Fish/Video\ at\ C_k) \quad (7a)$$

$$= \sum_{j=1}^{N_c} Var(Fish/Video\ at\ C_j) + 2 \sum_{j=1}^{N_c} \sum_{k>j}^{N_c} Cov(Fish/Video\ at\ C_j,\ Fish/Video\ at\ C_k) \quad (7b)$$

Equation 8: Measure of variance in fish abundance observed from several cameras, i.e., the variance of equation (7) results. $Fish/Video\ at\ C_j$ is the *Mean Fish Count per Video* for camera j, i.e., the result of equation (4) for one camera. $Var(Fish/Video\ at\ C_j)$ is the variance of $Fish/Video\ at\ C_j$, i.e., the result of equation (5) for one camera. $Cov(Fish/Video\ at\ C_j,\ Fish/Video\ at\ C_k)$ is the covariance of the results of equation (5) for cameras j and k. N_c is the number of cameras.

$$Cov(Fish/Video\ at\ C_j,\ Fish/Video\ at\ C_k) = \sum_{t=1}^{N_{10min}} \frac{(N_{t,j} - Fish/Video\ at\ C_j)(N_{t,k} - Fish/Video\ at\ C_k)}{N_{10min}} \quad (8)$$

Equation 9: Measure of covariance as used in equation (8). $Fish/Video\ at\ C_j$ is the *Mean Fish Count per Video* for camera j, i.e., the result of equation (4) for one camera. N_{10min} is the number of 10-min time periods covered by the video samples. $N_{t,j}$ is the number of fish observed at Camera j during the 10-min time period t.

To conclude, equations (4,5) provide relatively simple measures of fish abundance which overcome the issues of *Sampling Coverage* and *Fragmentary Processing*. However, these are applicable to analyze fish counts drawn from one single camera. For analyzing the overall fish abundance for several cameras, the applicable measure of fish abundance is given by equation (7). However, the related measure

of variance in fish abundance, i.e., equations (**??**), is not directly applicable in the case of missing samples due to *Sampling Coverage* and *Fragmentary Processing*. In such cases alternative methods exists [**?**] and can be chosen depending on each use case.

5 Visualizing Uncertainty Due to In-Situ System Deployment

Although the state-of-the-art does not offer well-established methods for quantifying the effect of *Fields of View* on fish counts, we provide users with elementary means to investigate their impact. A tab of the user interface is dedicated to the browsing of video samples, and is shown in Figure 8. Ecologists can inspect the different *Fields of View* over cameras and time periods. They can estimate which ecosystem is observed, which species are likely to be over- or under-estimated, and the potential *Duplicated Individuals*. Users can also investigate potential changes of field of view over time.

The lower part of the interface contains filtering widgets for selecting the videos of interest. Users can specify the characteristics of the videos of interest (e.g., time, location, *Image Quality*, species observed), in widgets that can be opened and closed on-demand. The widgets also offer an overview of the numbers of video samples for each characteristics. For instance, in Figure 8 the histograms represent numbers of videos over locations, year, and *Image Quality*. This offers basic means to investigate uncertainty due to *Sampling Coverage, Fragmentary Processing* and *Image Quality*.

Uncertainty due to these factors can be further detailed in another tab of the interface, shown in Figure 9. This tab offers the same widgets, and overview of numbers of video samples. Numbers of videos can be detailed in the main graph, on the upper part of the interface. Further, the main graph and the widgets' histograms can also display absolute numbers of fish, and mean abundance per 10-min as in (7). Figures 9-10 show visualizations of these metrics. The main graph can also display boxplots for visualizing the variance of fish abundance over sets of samples.

6 Uncertainty Due to Both Computer Vision Algorithms and In-Situ Deployment

Ecologists need to evaluate the uncertainty in end-results. These are impacted by uncertainty factors due to both computer vision algorithms and system deployment. Uncertainty factors interact with each other, as summarized in Figure 1. Although there is a variety of factors and interactions between them, their overall impact can be synthesized as two types of effect: noise, i.e., random errors yielding measurement variance, and biases, i.e., systematic errors yield under- or over-estimated measurements. Biases occur when measurements are systematically different under con-

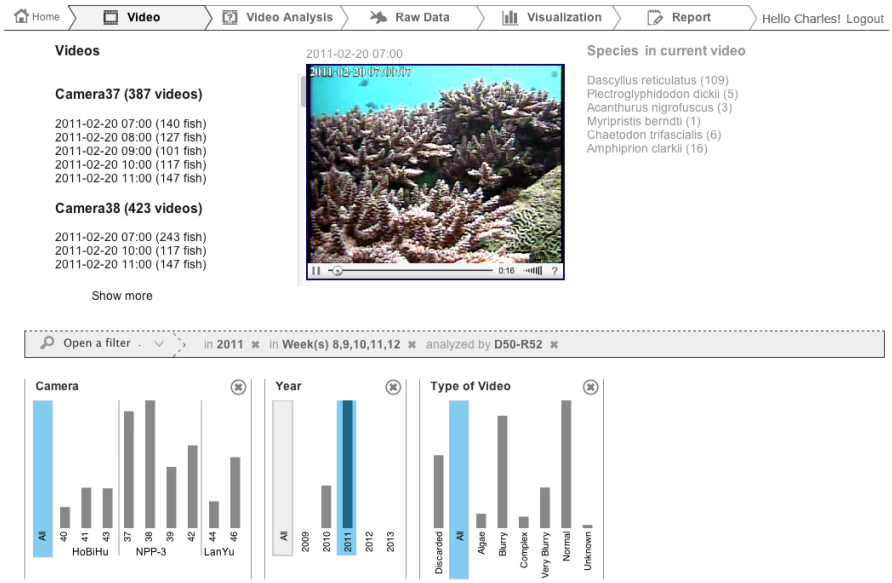

Fig. 8: **Video tab:** Video browser and visualizations for estimating uncertainty due to *Field of View*, *Duplicated Individuals*, *Image Quality*, *Sampling Coverage* and *Fragmentary Processing*. The bottom histograms show numbers of 10-minute video samples, and their distribution over locations (e.g., cameras), time (e.g., year) and *Image Quality*.

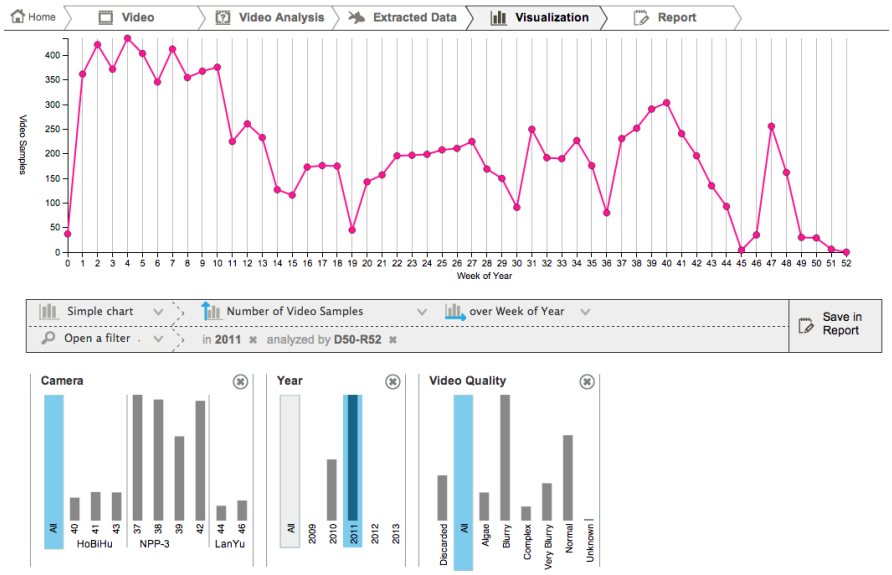

Fig. 9: **Visualization tab:** Visualizations for estimating uncertainty due to *Sampling Coverage*, *Fragmentary Processing* and *Image Quality*. The bottom histograms are the same as Fig. 8, and the main line graph above details the distribution of 10-minute video samples over one year (2011).

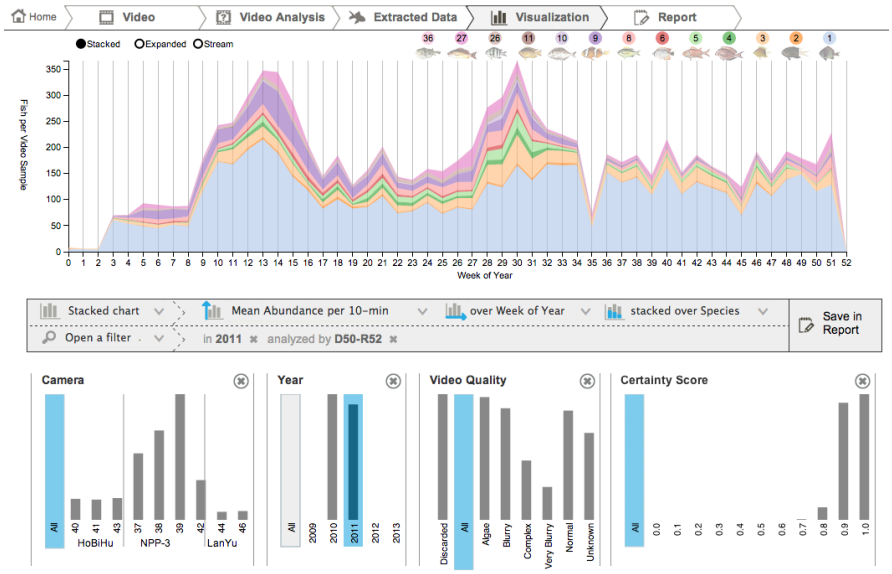

Fig. 10: **Visualization tab:** Visualizations for estimating *Uncertainty in Specific Output* due to *Sampling Coverage*, *Fragmentary Processing*, *Image Quality*, *Fish Detection Errors* and *Species Recognition Errors*. The bottom histograms and the main graph above show average fish counts per video, a balanced metric of fish abundance addressing *Fragmentary Processing* issues. The Video Quality widget shows fish abundance for each *Image Quality*. It indicates potential biases due to *Fish Detection Errors* that can arbitrarily vary depending on *Image Quality*, rather than natural phenomena. The Certainty Score widget shows the distribution of fish *scores*, which are used by computer vision algorithms to represent the similarity between each fish and their species models (section 2). These indicate potential biases due to *Species Recognition Errors* since errors are more likely to occur for fish with low score.

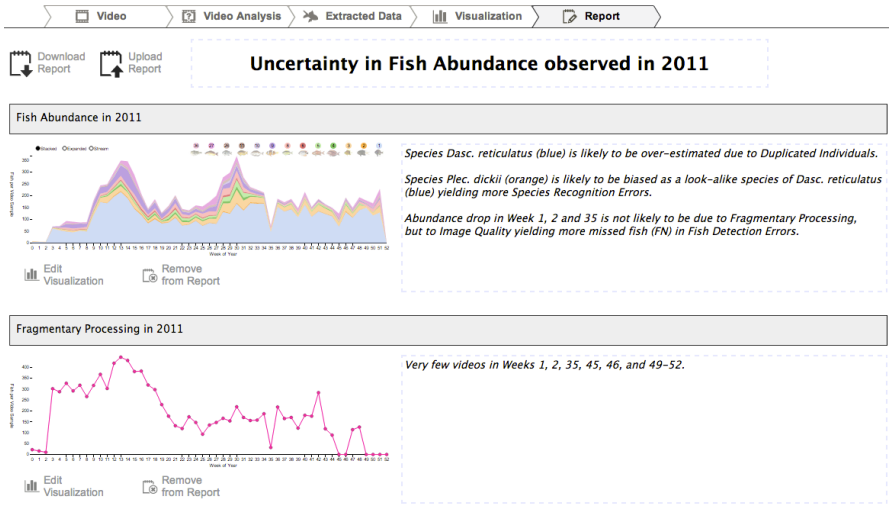

Fig. 11: **Report tab:** Example of visualizations gathered and annotated for describing *Uncertainty in Specific Output* due to *Noise and Biases*.

ditions that are independent from natural phenomena, such as *Image Quality*. This section discusses the means to measure the level of noise in the data, and to identify systematic differences of measurements. Levels of errors are usually measured under controlled conditions, e.g., using ground-truth datasets with specific characteristics. This section also presents the means to investigate errors in specific sets of end-results, which characteristics can be different than those of ground-truth sets.

Noise and Biases - Random errors, i.e., noise in end-results, is commonly measured using metrics of mean and variance such as (4,5). These metrics are a well-established basis for statistics investigating of all sorts of populations [**?**]. Significant differences of means and variances can be observed under conditions that are independent of natural phenomena. An example can be considered with a same fish population observed with different image qualities, or fields of view. If computer vision yields significantly different means and variances of the fish counts, then the different observation conditions can potentially bias the end-results.

As mentioned in section 4, no well-established methods are available for evaluating biases due to *Field of View* and *Duplicated Individuals*. Hence the rest of the discussion focuses on identifying biases due computer vision algorithms. As discussed in section 2, the Fish4Knowledge project was able to measure *Fish Detection Errors* for each *Image Quality*, and *Species Recognition Errors* for each species. Such measurements can support the evaluation of potential biases due to *Image Quality* and *look-alike* species.

If measurements of *Fish Detection Errors* (e.g., equations (1,2) and Figure 3) vary significantly with different *Image Quality*, then they indicate potential biases in end-results. Sets of end-results from a specific image quality can be artificially over- or under-estimated, compared to end-results from another image quality. If error rates are of the same magnitude for all image qualities, then they do not indicate potential biases. They rather indicate a general level of noise, even if error rates are high. End-results drawn from image qualities having similar levels of uncertainty are potentially over- or under-estimated in the same way, and hence, are comparable. Contrarily, high error rates for *Species Recognition Errors* indicate potential biases between look-alike species, even if error rates are of the same magnitude for all species.

The visualizations of *Fish Detection* and *Species Recognition Errors*, presented in section 3.3 and Figures 3-5, support the identification of significant difference in error rates indicating potential biases in end-results. We assume that the significance of error magnitude depends on the study at hand, and their specific requirements with uncertainty issues. For instance, a descriptive survey of fish population may tolerate higher uncertainty than a survey intended to demonstrate causal effects of specific environmental conditions.

Uncertainty in Specific Output - Measurements of *Fish Detection* and *Species Recognition Errors* in ground-truth sets potentially support extrapolations of errors in other sets of computer vision results. Error rates in equations (1-3) can be used to extrapolate errors in end-results, by multiplying them with the numbers of fish

in the output. For instance, given a set of fish detected in video samples of image quality Q_i, the potential number of False Positives in end-results could be computed using equation (10).

$$Non\text{–}Fish\ in\ Samples\ Q_i = N_i\ *\ Type\ I\ Error\ Rate\ Q_i = N_i\ \frac{FP_i}{TP_i + FP_i} \tag{9}$$

Equation 10 *Non–Fish in Samples* Q_i is the extrapolated number of False Positives in a set of end-results extracted from videos of image quality Q_i. N_i is the number of fish in the end-results. TP_i and FP_i are the numbers of True Positives and False Positives measured for a ground-truth set of image quality Q_i.

However, the validity of such extrapolation relies on the assumption that errors measured in ground-truth evaluations are representative of errors occurring in computer vision outputs. Further research is needed to control this assumption. For instance, the proportions of fish and non-fish objects may vary across videos of the same image quality, and this would bias the results of (10). Alternative methods exist for the case of varying class proportions [**?**], and can potentially provide more accurate counts of individuals. However, future work is needed to assess their reliability. Hence the Fish4Knowledge user interface did not retain uncertainty methods such as (10). Metrics such as (1-3), complemented with numbers of fish and videos samples in sets of end-results, were retained for simple indications of uncertainty in end-results, without extrapolating the numbers of errors.

To complement the evaluation of uncertainty in end-results, the user interface can display the *certainty scores* measuring the resemblance of each fish with the model of its species. Figure 10 shows the widget conveying the distribution of fish over *certainty scores*. Species models are constructed using ground-truth images dedicated to the learning of fish appearances. *Scores* are used by the *Species Recognition* algorithm for selecting which fish to classify in each species. The higher the *score*, the more likely the fish truly belongs to the species, and the lower the chances of errors. The *score* is not a measure of error probability, but a measure of visual similarity between fish occurrences and fish models. Measures of error probability can be developed on the basis of this *score*, and such probability can be used to improve the computation of fish abundance (see Chapter **??**).

We investigated the impact of such *scores* on user understanding of uncertainty [**?**]. As shown in section 3.2, user trust and acceptance was slightly improved by providing *score* thresholds to select fish to retain in end-results. Hence we retained the use of such *score* in the user interface. A filter widget displays the distribution of fish over *scores*, and allows the manual selection of a threshold (see Figure 10).

7 Future Work

Ground-truth evaluations are well-established methods for evaluating uncertainty due to computer vision algorithms. However, future work is needed to enable the extrapolation of errors in end-results. The representativity of ground-truth needs to be assessed. Large numbers of ground-truth items are selected amongst the entire collection of images samples. This ensures *a priori* that the ground-truth is representative of the entire video collection. However, this method does not demonstrate that the magnitude of errors measured for the ground-truth sets is similar to that of computer vision performed on other video sets. An approach to estimate how the ground-truth is generalizable could consist of repeating ground-truth measurements, and computing the mean and variance of numbers and rates of error. This method can support extrapolation of errors in end-results, and the measure of confidence intervals for extrapolated errors. But it may require an extensive ground-truth collection. Another approach can make use of error probabilities estimated from *certainty scores*, as discussed in Chapter **??**. The accuracy and the costs of these approaches can be compared.

During user interviews, ecologists often asked how to evaluate if the ground-truth sets contain enough fish examples. This aims at estimating the cost of ground-truth collection implied for integrating the detection of a new species. It also aims at deciding on collecting further ground-truth items for the species that are difficult to recognize. Future work is needed for establishing methods to estimate optimal ground-truth size. An approach could consist of repeated ground-truth evaluation for the same computer vision algorithm, but trained using ground-truth sets of different sizes. If the numbers of errors are relatively stable although ground-truth size is increased, then users can consider that the number of ground-truth items is sufficient.

Uncertainty due to in-situ deployment of the system requires important future work. Metrics for *Duplicated Individuals* need to be researched, and to take into account the species and *Fields of View* at stake. Such metrics can be of the same form as $Type\ I\ Error\ Rates$ in (1), and *Duplicated Individuals* can be considered as False Positives. To extrapolate *Duplicated Individuals* in end-results, the measurements can be repeated over ground-truth sets to compute the mean and variance. Similarly to extrapolations of computer vision errors in end-results, this supports the estimation of confidence intervals for the numbers of *Duplicated Individuals* extrapolated in end-results. Finally, future work needs to address the challenge of extrapolating potential biases and errors in end-results by taking into account the different uncertainty factors. To do so, a unified framework of compatible uncertainty metrics needs to be researched. It needs to integrate metrics of biases and errors from *Fish Detection* and *Species Recognition Errors*, *Image Quality, Fields of View*, and *Duplicated Individuals*, and metrics of species abundance accounting for *Fragmentary Processing* and geo-temporal *Sampling Coverage*, e.g., average fish counts per unit of time or area.