Technische Universiteit
Eindhoven
University of Technology

Video Analysis Software for Scientific Use: Impact of Data Provenance Information on User Trust, Acceptance and Information Processing

Elvira Arslanova

978-90-444-1230-7

Acknowledgements

I express my deepest appreciation to my external supervisor *Lynda Hardman* for her expertise and knowledge that she shared with me during the whole process. She provided enough freedom of choice in research direction and support in every phase of the process. I am grateful to Lynda for her patience, availability to answer my questions and interest in all the details of the working process. Besides all, her inspiring personality on both professional and a personal level guided me in my work.

I would like to thank my internal supervisor *Rene Ahn* for his advice and knowledge in technical aspects of the project. His feedback encouraged and challenged me during the working process.

My work would have been impossible without the help of my colleague *Emmanuelle Beauxis-Aussalet.* Thank you for interesting and fruitful discussions, sharing your ideas and experience. Emmanuelle contributed a lot to enhancement of both my professional and teamwork skills.

I would like to thank *Tiziano Perrucci, Jiyin He, Jacco van Ossenbruggen* and all the members of the *Fish4Knowledge* team for their technical support.

I thank the *biologists* who valuably contributed to the project by participating in our studies. I specially thank *Tony Fan* and his colleagues from the Coral Laboratory in National Museum of Marine Biology and Aquarium for collaboration and help provided us during our trip to Taiwan.

Lastly, I would like to thank my *family*, *friends* and my *partner* for the great support and encouragement during my studies.

Summary

We present a case study of scientific usage of video analysis software for monitoring fish populations in the context of the EU Fish4Knowledge project. While video analysis systems are not yet widely used in the marine biology domain, analysis of captured videos is becoming a wide-spread technique to capture data on marine environments. Fish4Knowledge provides long-term video monitoring of fish populations using underwater cameras. Automated video analysis introduces uncertainty in interpretations of sightings of fish used by marine biologists and influences their acceptance of the software. We investigated how user acceptance, trust and understanding are impacted by providing data provenance information on the video analysis process.

Including data provenance based on video analysis evaluation information requires significant mental effort to understand in the context of marine research. We discuss possible information processing issues, for example attention narrowing, goals achievement and working memory loss, that could be faced while using the data analysis tools in the project.

Following the user research findings, we propose guidelines for designing the information required and interfaces to access it in data analysis tools based on video analysis.

Executive Summary

This report presents the findings of a User System Interaction research project as a part of the Fish4Knowledge project. We explore design issues related to video analysis software targeted to marine biology researchers.
We investigate whether data provenance influences users' trust in the Fish4Knowledge tool (Section 2). We conducted a study to find the dependencies between the amount of technical details revealed to users and their trust in competence of the tool, acceptance and understanding of the underlying processes. We also determined data provenance needs based on interviews with users. The study showed that

- users require data provenance information to decide on the appropriate statistical method for data analysis;
- user trust is not influenced by the depth of technical details;
- acceptance of the automated video analysis tools in the marine biology community is high in spite of the uncertainties inherent to the technique;
- users require information about camera settings, habitat description and performance of the software in different conditions.

We evaluated the extent to which the Fish4Knowledge tool supports information processing (Section 3). We adapted a situation awareness framework to measure user performance depending on the complexity of questions. We highlighted issues related to goal achievement, human attention, working memory and expectations that could be faced while using the Fish4Knowledge tool. The cues provided by the Fish4Knowledge tool did not guide users to check the data provenance. While using complex data analysis systems such as Fish4Knowledge, users encounter issues with attention tunnelling and do not pay attention to the parameters of the analysed dataset.
Based on the studies we carried out, we determined the list of design recommendations relevant to automated video analysis in scientific environments (Section 4):

- Provide data provenance information.
- Give access to ground truth collection.
- Verify the algorithms by fellow biologists.
- Verify the automated method using accepted data collection methods.
- Provide performance indicators of the video analysis in different conditions.
- Place salience on the parameters of the dataset.
- Provide explicit cues to verify the validity of data visualizations.

Contents

1 Introduction

A method to acquire observational data of insects, fish and animals in their natural setting is to use embedded video cameras to observe presence and behaviour of the organisms. Environmental videos could help to diagnose the harmful impact of global warming, pollution or natural disasters. The Fish4Knowledge project has been continuously recording video footage of coral reef fish from a number of underwater cameras since 2010. The video streams are being generated as part of an ecological monitoring effort and form a resource base for marine biology researchers. However, extracting useful information from such quantities of data requires an overwhelming number of hours of human effort to analyse the collected videos. An alternative to costly human analysis is to use automated software tools. The Fish4Knowledge project investigates methods for how a combination of computer vision, database and workflow methods can be used to extract useful information and make it accessible to non-programming biologists.

Computer vision techniques introduce new types of uncertainty and errors to the marine biology field (e.g., certainty of species recognition), which could result in distrust of the output and lack of acceptance from users. Providing data provenance information could increase confidence in the results and acceptance of the tool by the community. However, the technical complexity of the provenance information could lead to misunderstandings of the concepts and further confusion. While marine biologists are not experts in video processing tools and their technical concepts, they need to have some minimal understanding of the limitations of them to be able to use them in their research. In order to provide an environment that allows biologists to accept the results of the automated system we need to understand their requirements for interpreting the analysis results . In particular, we need to:

- determine data provenance information necessary for users;
- understand the potential impact of introducing data provenance on *user trust* and *acceptance*;
- determine the required level of *understanding* of the underlying video analysis processes;
- compare the limitations of video analysis compared to other data collection methods.s

Our insights are also relevant for solving trust issues encountered with similar video analysis tools delivering provenance information for scientific video collections in other fields.

While interacting with such complex systems as Fish4Knowledge, users could become overwhelmed with complexity of their task and overloaded by data. This could cause mistakes in their working process and, thus, conclusions. For instance, biologists could confuse the real biological phenomena and an artefact caused by the analysis or sampling method. To avoid this, the interface should provide sufficient cues to be able to effectively direct their attention to salient aspects of the data (e.g. fish counts, performance of software, quality of data). These should be intuitive to guide them in their exploration and satisfy their current expectations how to handle the limitations of the analysis. While processing information, users could also experience issues with achieving their goals due to the working memory overload or inactivated sub-goals (e.g. check validity of the fish counts). To prevent these mistakes, it is important to provide good situation awareness of the processes happening with the data. We adapted the concept of situation awareness from Endsley [5] to evaluate the extent to which the Fish4Knowledge interface directs users in their analysis.

Fish4Knowledge is a data analysis tool, which makes it different from typical domains where situation awareness measurements are applied. While time and space are dominant in situation awareness, they are not important in our domain. However, the risk of data overload and complexity of the underlying processes that are poorly understood requires sufficient situation awareness. It does not guarantee that the best decision would be made. However, without it, a good decision is only a matter of luck. We evaluate whether the Fish4Knowledge interface provides sufficient cues for biologists to be aware of all the processes happening with the dataset of interest without overwhelming them. In the context of Fish4Knowledge, we investigate the following questions:

- the potential issues related to *achieving goals*, human *attention*, *working memory* and *expectations*;
- the dependency between the task complexity and human performance.

Fish4Knowledge is tailored to marine biologists with research interests in coral reef fisheries and aquatic ecology.
Fisheries scientists study systematics of marine fishes, fish community and population dynamics. They traditionally use diving observations to collect data of fish counts and behaviour. To model ecosystems biologists require such metrics as fish abundance, distribution and biodiversity. The Fish4Knowledge tool could provide information about detected species and, potentially, the fish behaviour.
Aquatic toxicologists study the effects of manufactured chemicals and other anthropogenic and natural materials and activities on aquatic organisms at various levels of organization.The Fish4Knowledge tool could support their studies on monitoring fish population at sites with different contamination levels.

Requirements supplied by the Fish4Knowledge project Based on interviews to obtain user requirements, the Centrum Wiskunde & Informatica (CWI) team summarized these as the 5 requirements defined below [2, 3].

A. *Support the analysis of population dynamics* by providing the following metrics: abundance, species composition and species richness.

B. *Support browsing videos of interest*, in particular those that correspond to the datasets that are visualized.

C. *Support the identification and the correlation of trends* in selected datasets and metrics of interest. For instance, the feeding time of a particular species or heir highest activity time.

D. *Provide an overview of the uncertainty inherent to each video analysis component.* This overview must report on i) the ground truth dataset, its inherent errors due to involvement of human judgement; ii) the related machine learning evaluation of the components; and iii) the certainty score profile of the components derived from the analysis.

E. *Provide an estimation of the potential errors contained in visualized datasets.* This estimation of errors can be expressed using 2 types of inter-related metrics: i) the certainty scores, and ii) the estimated numbers of True Positives, False Positives and False Negatives. From a high- level point of view, these are intended for users to evaluate the level of confidence in the trends, and correlations of trends, they identify.

The CWI team developed a first high-fidelity version of the Fish4Knowledge prototype [3] and after that created a second iteration of improved paper mockups. This implementation was used to collect early user feedback and explore the means to develop the web-interface. We explored the means to address the requirements described above. To discover the user needs in information about inherent uncertainty (*Requirement D*) and potential errors (*Requirement E*), we conducted a user study on data provenance needs and their impact on user trust (Section 2). In our interface we support the analysis of population dynamics (*Requirement A*) and identification of trends (*Requirement C*). To evaluate the extent to which we meet these requirements and whether we provide sufficient cues to guide the analysis, we conducted a second user study described in Section 3. *Requirement B* is not covered in this report and requires further investigation.

2 Data Provenance Requirements and Their Impact on User Trust

Video analysis is a novel technique in marine biology and researchers need to be confident in scientific validity of the results produced by the Fish4Knowledge tool. Moreover, they need to be aware of inherent uncertainty and errors possible in the data to be able to compensate for them in their studies. In cases when research is done on data collected by others, a common approach to establish trust in information resources is to present the origins of data [4, 14, 16]. The data provenance concerns the systematic recording of the derivation history of each data item, starting from its original sources. These records are often of crucial importance for a researcher to decide if the data is suitable.Tool support for maintaining data provenance is common in scientific workflow systems such as Kepler [15] and Taverna [20].
Indicating data origins could potentially improve user trust in the software. A number of models and scales were used to measure Human-Computer Trust (HCT) in different computational systems [1, 21]. While research in trust in information resources has a number of valuable contributions, the majority of them are relevant only for web-semantics or internet domains. Madsen proposes a psychometric instrument specifically designed to measure trust in computer systems. We adapt his definition of trust [17, 19]:

Trust is the extent to which a user is confident in, and willing to use the data produced by the video analysis software.

Madsen determines 4 *aspects of trust* in software [17]:

- user trust in the technical competence of the video analysis;
- user acceptance of the video analysis;
- user understanding of the technical concepts used to create the analysis;
- user satisfaction of the information need for data provenance.

Confidence and willingness to use software are influenced by cognition and affect-based components of HCT. Specifically, cognitive based components of HCT are impacted by the explanations provided by the software [18, 24].
An explanation of the underlying processes could influence user trust by increasing confidence in technical competence of the software, acceptance. However, technical details of computer vision technology can be difficult to understand for people without technical background. To our knowledge, no research has been done in determining what and how much data provenance about video analysis should be revealed to non-technical scientists. In our study we explore the depth of details that should be shown to users and their impact on user trust in the competence of the software and acceptance. We investigate whether revealing data provenance improves user understanding and satisfies his/her information needs.

2.1 Method

2.1.1 Setup

To investigate methods of coping with uncertainty and compare the limitations of video analysis to another methods of data collection, we conducted semi-structured interviews prior to an experiment. We asked participants about their topics of research, information needs, their current data collection practices, corresponding potential biases, and their envisaged usage of our video analysis tool. To measure user trust we adapted a questionnaire from [17].

To define the amount of provenance information that should be provided, we designed three mid-fidelity prototypes (Fig.1, 3, 5). The prototypes were gradually increasing amounts of information about the underlying processes. In each prototype we explain the evaluation of fish detection and species recognition components using the comparison of automatic detections with manual detections from a ground truth collection. Each subsequent version contains the information explained in the previous one (e.g. interface 2 reveals more details about the information presented in interface 1).

All three interfaces have the same layout and consist of *Video*, *Video Analysis* and *Visualization* tabs. The *Video* tab allows users to browse the videos of interest. The *Visualization* tab illustrates the fish counts - a predefined static chart based on the data contributed to the experiment. The *Video Analysis* tab explains the ground truth evaluation of the Fish4Knowledge analysis tools.

We introduced the basic technical concepts which are necessary to understand the measure in 3 subsequent versions (see Table 2.1.1 for more details). The measure was presented as the comparison of each fish image with a fish model. The provenance information provided a set of fish counts produced by using different thresholds on the feature vectors.

Interace	Explained concepts:	Picture
1	• The concept ground truth. • The comparison of manual and automatic fish counts. • The number of videos in the ground truth dataset and in the large overall collection.	1, 2
2	• All points in the interface 1; • The machine learning evaluation: True Positive, False Negative and False Positive.	3, 4
3	• All points in the interface 2; • Feature vectors and similarity thresholds of classifiers.	5, 6

Table 1: Description of interaces used for the experiment

2.1.2 Participants

We recruited 10 participants (1 professor, 8 researchers, 1 master's student) working in the biology research domain from the Netherlands and Taiwan using the snowball technique. We recruited participants working in fisheries domain and studying fish population dynamics. The participants belonged to one of the following groups of biologists:

Deep sea fishery. Several teams of deep sea fishery researchers from the Dutch institutions participated in our studies. They study fish abundance, distribution and biodiversity. These teams traditionally use commercial or experimental fishing to sample data. One team records the video moving on a fishery vessel with a lighted cameras attached and calibrated on a specific distance from the sea bed.

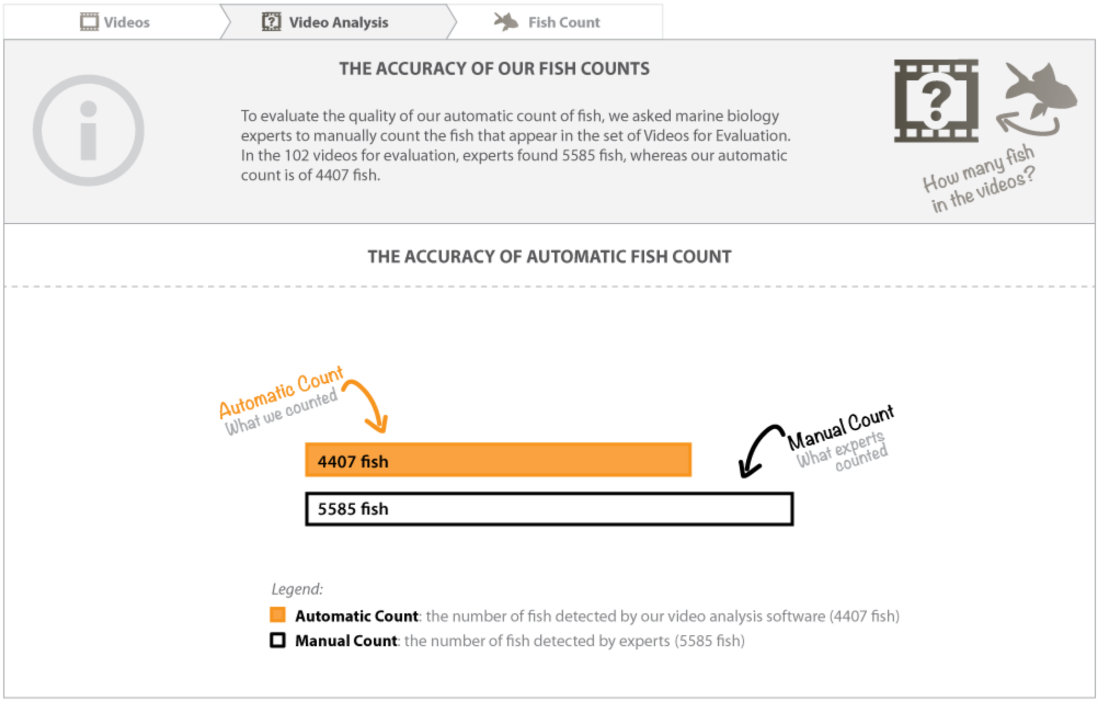


Figure 1: The Video Analysis tab with minimum amount of details

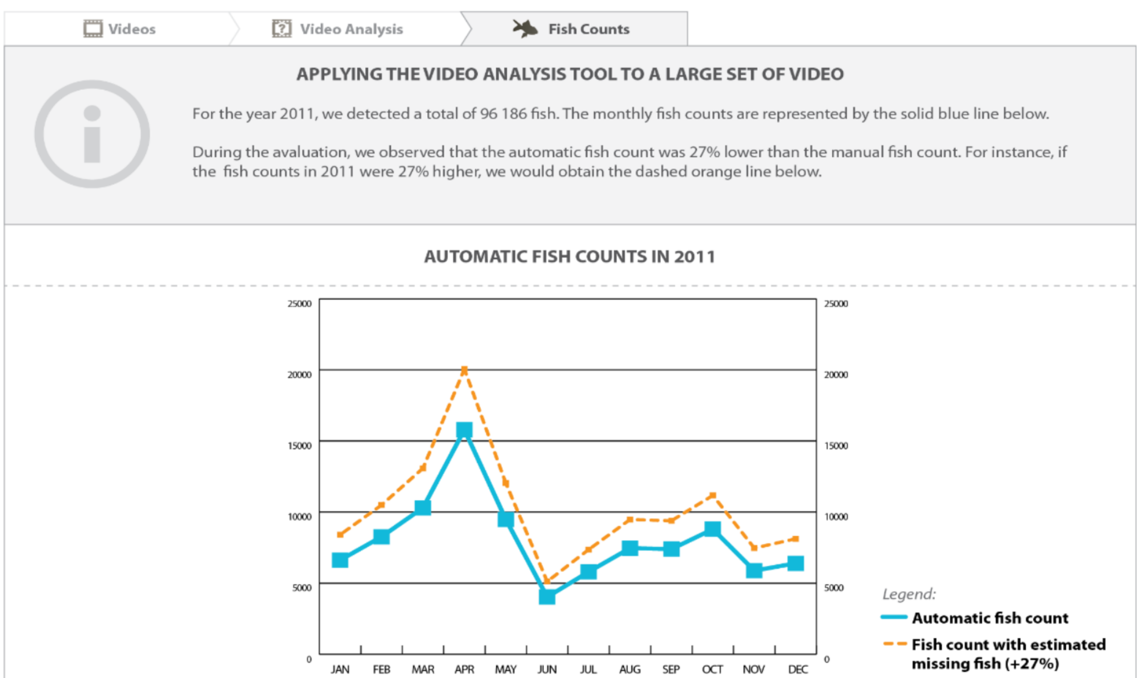


Figure 2: The Visualization tab with minimum amount of details

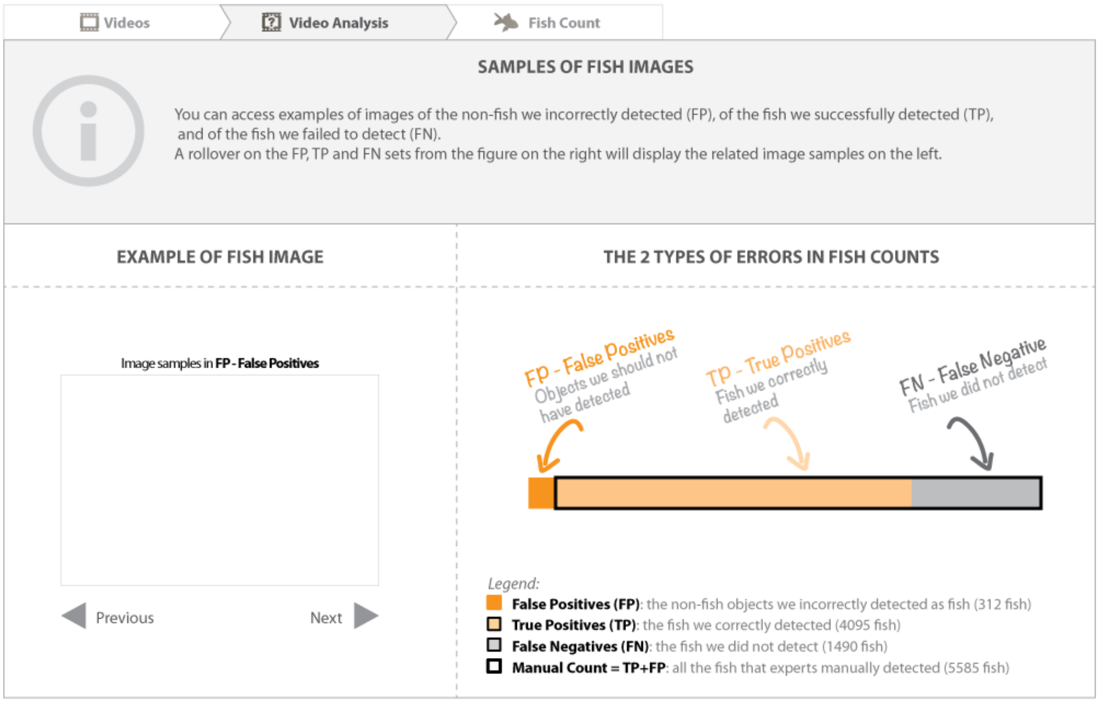


Figure 3: The Video Analysis tab with the average amount of details

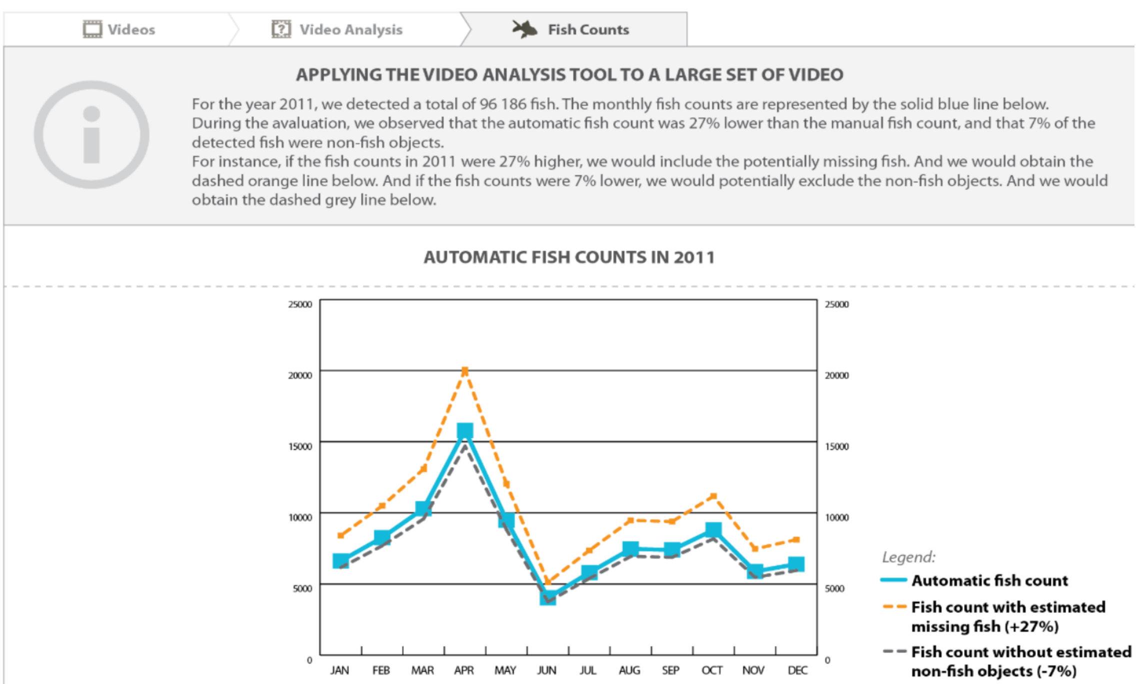


Figure 4: The Visualization tab with the average amount of details

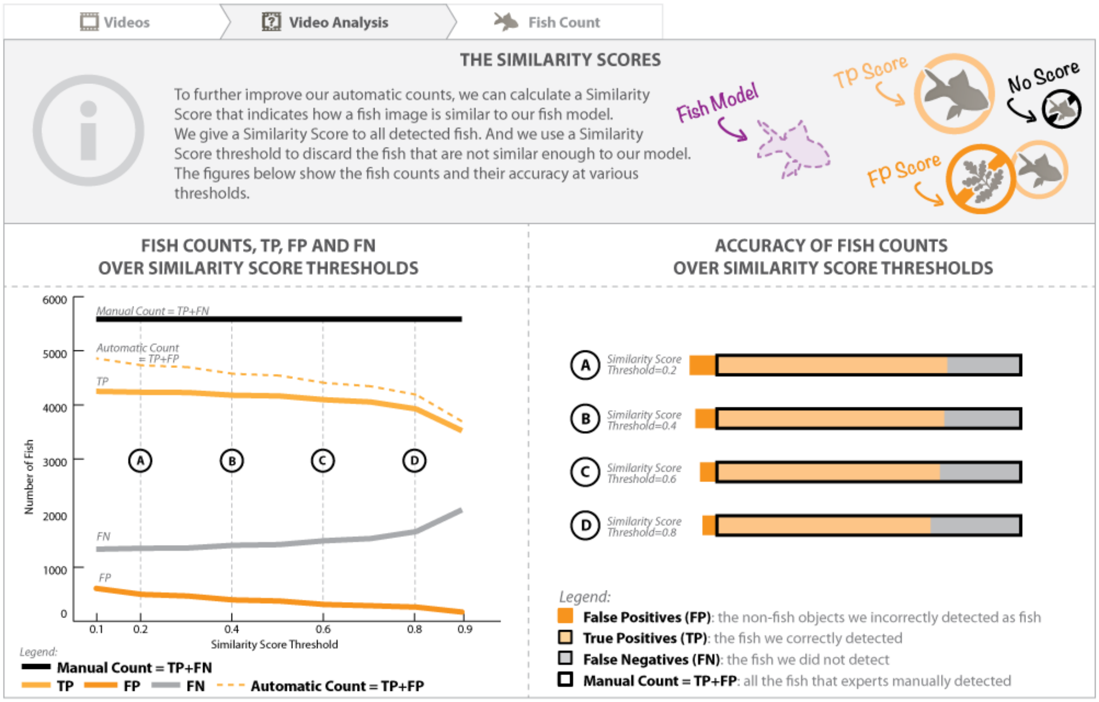


Figure 5: The most detailed Video Analysis tab

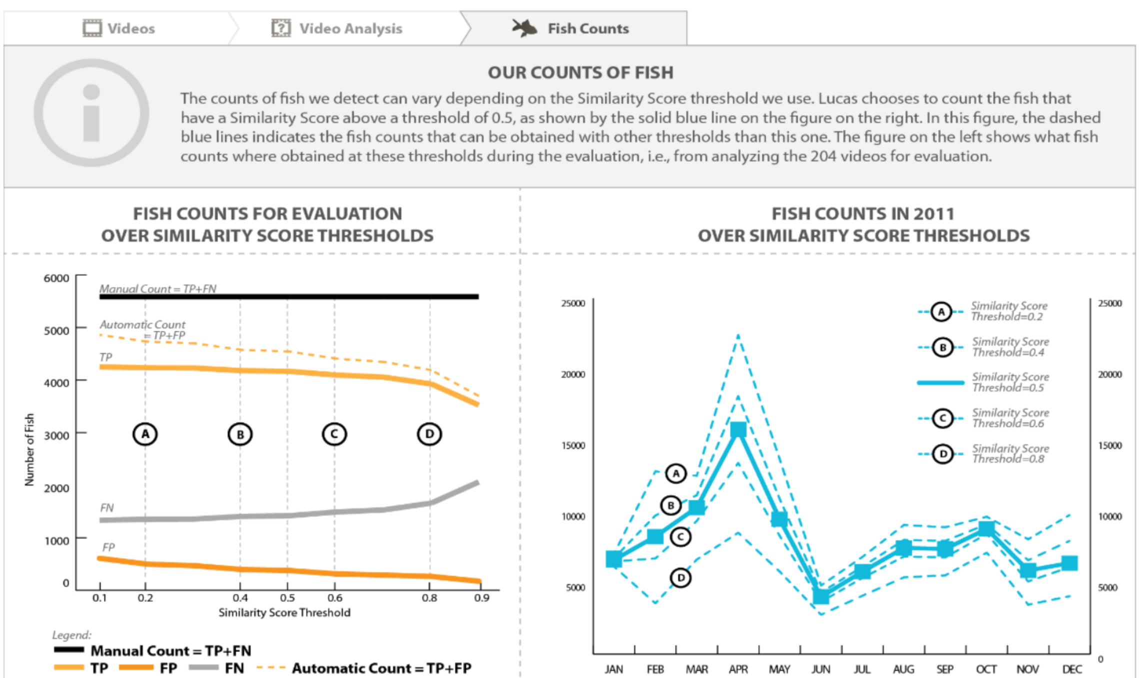


Figure 6: The most detailed Visualization tab

Coral reef fishery. Biologists who study population dynamics and fish systematics of coral reef fish. They traditionally use diving observations to collect data. Some teams already use video cameras to collect data for researching fish systematics, morphology and evolution.

2.1.3 Procedure

We conducted a semi-structured interview prior to exposing users to the interface. After that the participants were given initial instructions in written and oral forms. Depending on the geographical and personal availability of the participants they accomplished the tasks in the presence of the interviewers or remotely via internet. Four participants were not able to complete the experiment, but participated in the interview.
The participants who completed the experiment in presence of the interviewers, were trained to use the *Think Aloud* method on an independent web-site. They were asked to assume the role of a biologist interested in trends in fish populations during 2011 and asked to examine the interface. After each interface version, users answered a questionnaire about the technical concepts to measure the understanding level and highlight the potential issues. Lastly, they were asked questions that measured their trust in the tool, the acceptance of the tool's imperfection, and the satisfaction of their information needs.

2.1.4 Analysis

The user interviews were transcribed and independently interpreted by two researchers in the Fish4Knowledge team. We focused on the comparison of implicit biases associated with the different data collection methods. This is useful to understand how users are likely to deal with the potential bias introduced by video analysis.
Using the *Affinity Diagrams* method, we classified the user feedback (interviews and comments made while *thinking aloud*) into 4 categories: (potential) acceptance of the tool, understanding of the underlying processes, technical competence of the software and user information needs. We analysed the answers on technical questions.
Although the low number of participants is not representative of the marine biology community, we conducted the quantitative analysis of user technical competence, acceptance and information needs to compare the results of the qualitative study.

2.2 Results

2.2.1 Results of the Qualitative Analysis

Comparison to other data collection methods Based on the user interviews, we synthesized information on current data collection methods, uncertainty issues and potential of video analysis (Table 2.2.1). The uncertainty introduced by video analysis is similar to their usual data collection methods' uncertainty. Their widely-accepted and commonly-used methods can contain significant biases, such as the under-estimation of some species (e.g., during diving observations divers could scare some species of fish). Biologists usually cope with these uncertainties by applying robust statistical methods, by collecting extensive numbers of samples, and correlating different data sources. Biologists usually refer to this acceptance of uncertainty as a trade-off between precision (small magnitude of errors, e.g., small error bar) and accuracy (closeness of measurements to the correct feature, e.g., the right set of species and habitats).

Data	Data Collection	Sampling Method	Uncertainty Issues	Research Topic	Potential of video analysis
Fish count, Species, Size	Video Images:baited stereoscopic camera, manual analysis.	Singlepoint locations	Avoid duplicated detection of single fish Few overlaps.	Population dynamics, Migraton	To avoid manual image analysis in current practices
Fish count, Species, Size, other objects	Video Images: lighted camera held close to deep sea floor, calibration of distance to seabed, manual analysis	Transects (along a virtual line)	Rare misidentification of species. Cryptic organisms.	Population dynamics, Trophic systems	To avoid manual image analysis in current practices.To reduce expensive vessel usage.
Fish count, Species, Size	Diving observations, handheld camera for backup purposes	Transects	Diversity of ecosystems. Some species are hiding from divers.	Population dynamics, Trophic System	To analyze existing videos. To avoid diving
Fish count, Species, Size, Weight, Bone size, Chemicals, Nutrients [1]	Experimental Fishery with fish dissection	Singlepoint locations or transects	Variability of fish catch in same conditions.	Population dynamics, Trophic System, Migration, Reproduction	Excluded due to unsupported or imprecise measurements
Fish count, Species, Size, Weight	Commercial Fishery: data for North-Sea fish market	Dependent on commercial fisheries	Variability of fish catch. Only commercial species are targeted. Uncommon species are misidentified.	Population dynamics, Migration, Reproduction	To compensate the biases in the market-dependent sampling
Fish count, Species, Size	Diving Observations	Singlepoint locations or transects	Diversity of ecosystems. Some species are hiding from divers.	Population dynamics, Trophic systems, Reproduction, Environmental event	To avoid diving.
Fish count, Species	Video Images & Commercial Fishery: video analysis of fish discarded during on-board fish processing.	Dependent on equipped commercial vessels	Misidentification of fish and non-fish objects	Population dynamics	Currently experimented, needs improvement.

Table 2: Results of the Interviews Grouped by a Research Team

Technical Competence Most participants did not trust the technical competence of the tool. The details introduced in Version 2, Fig. 3, did not improve user trust. In Version 3, however, user trust was improved in 4 cases out of 7, and worsened in 1 case. The most sceptical participants showed a very good understanding of the technical details, whereas the most confident participants showed a poor understanding .

Acceptance Most participants would accept the tool for scientific purposes, even though the biologists were aware of missing 27 % of the fish, which was stated in the interface. Fish4Knowledge is the only available tool that allows long-term direct observation of underwater ecosystems. The details introduced in Version 2 had no major impact, but in Version 3 acceptance was improved in 2 cases. All participants answered that amongst the 3 interfaces they would choose to work with the provenance information of the 3rd version. The availability of alternative fish counts for several thresholds was important for participants, as they can choose the most appropriate threshold for their case (e.g., "*I want you to give me as many lines* [lines representing thresholds of similarity score in the Version 3] *as possible and I will decide which one to use*"). All participants that were already using videos in their research showed a high level of acceptance.

Understanding At all steps, most participants had difficulties understanding the technical concepts in the presentation. We also observed many misunderstandings of the technical questions of the questionnaires. Participants needed a long time to think, to read the explanations several times, and requested for extra oral explanations. For example, most participants showed difficulties with recalling the definition of new terms we introduced (e.g., False Positives). Interface 1 was better understood but its basic concepts were not grasped by 2 participants. Two participants mentioned that they would ask the help of a fellow biologist, whose expertise lies in statistics, if the evaluation of the technical details is too complex or too long. The participants with an excellent understanding were already familiar with the technical concepts.

Information needs All participants expressed a strong need for more information, in all steps of the experiment. The mostly wanted information on aspects not covered by the given explanations:

- Description of the *sampling method* supports users in controlling that the species of interest can be observed in the collected data, namely the coverage of the video collection and of the extracted features.
- Description of the *potential errors* in video data supports users in *differentiating trends from noise in data*, and identifying meaningful patterns. It includes the machine learning evaluation of video analysis performance, which leads users to having to with technical concepts beyond their current expertise, such as False Positive, True Positive, False Alarm Rate, Detection Rate, features vectors and classifiers.
- Traditional statistical measurements in marine biology, such as *mean* and *standard deviation*, presented in a form of error bars.
- Access to the *ground truth collection* related to the visualization.
- *Camera settings*, such as location, camera field of view and description of a habitat, helps biologists to control the biases introduced by the factors like change of the field of view.
- Comparison of the evaluation to other data collection methods (e.g., diving observations) helps biologists without technical background assess the performance using familiar concepts.
- Details of video analysis performance under various conditions (e.g. water quality).

2.2.2 Results of the Quantitative Analysis

We analysed the impact of the different interfaces on perceived technical competence, acceptance and information needs.
We analyzed the data using a repeated measures ANOVA, with a Greenhouse-Geisser correction due to violations of sphericity (Mauchly's Test $\chi^2(2) = 6.596, p = 0.037$)(Fig. 2.2.2). A significant effect was found between the mean User Trust (Technical Competence) and different explanation types ($F(1.378, 16.541) = 3.86, p < 0.05$)(Fig.2.2.2). A Bonferroni-correction post-hoc test revealed that this effect can be attributed to the increase in User Trust (Technical Competence) from the condition with medium amount of information (Interface 2) in combination with the most extensive explanation (Interface 3) ($p = 0.004$).

2.3 Limitations

Most of the participants are not experts in coral reef ecosystems, which could potentially introduce differences in their perception of uncertainty (e.g. while working in murky waters they have to deal with different types of uncertainty than clean shallow waters).
Due to geographical limitations we did not interview and observe interactions of all the participants, which could have provided us with more insights about their responses on the questionnaires and tasks.
The sample size used for the quantitative analysis could be not representative.

2.4 Discussion & Conclusion

The Fish4Knowledge tool provides marine biologists with enormous amount of data. To our knowledge, this is the biggest collection of fish records that has ever been available. Marine biologists have statistical techniques to work with big data, however, they require data provenance information to choose the appropriate statistical method. Besides potential errors in data, users require information about sampling method, camera settings (e.g., field of view, location), varying image quality and evaluation of software under different conditions (e.g., for particular species, different water quality). Verification of the technically complex concepts could more easily be accepted by users if the comparison to the other methods of data collection (e.g., diving observations) is provided.
Most of the marine biologists did not have a technical background, and our study shows that they have rather poor understandings of the technical details (e.g., non-fish object detection, species misidentification). They do not fully trust the technical competence of the tool, however, our small-sampled quantitative analysis and qualitative data indicate that revealing more technical details (in our case, feature vectors and similarity thresholds of classifiers) could potentially improve trust in technical competence of the tool, but does not influence acceptance and understanding. The issues with understanding might have increased the needs in extra information, which were expressed by the participants.
Although understanding and trust in technical competence were low, the acceptance of the tool by the participants was high, both observations were supported in the interviews and questionnaires. This could partly be explained by the lack of direct observations of the environment offered by other data collection methods, whereas uncertainty is introduced by all of them (see Table 2.2.1).

Mauchly's Test of Sphericity[b]

Measure: User trust

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Explanation_Type	.549	6.596	2	.037	.689	.749	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

Tests of Within-Subjects Effects

Measure:User_trust

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Explanation_Type	Sphericity Assumed	2.000	2	1.000	3.866	.035
	Greenhouse-Geisser	2.000	1.378	1.451	3.866	.055
	Huynh-Feldt	2.000	1.499	1.334	3.866	.051
	Lower-bound	2.000	1.000	2.000	3.866	.073
Error(Explanation_Type)	Sphericity Assumed	6.208	24	.259		
	Greenhouse-Geisser	6.208	16.541	.375		
	Huynh-Feldt	6.208	17.985	.345		
	Lower-bound	6.208	12.000	.517		

Pairwise Comparisons

Measure:User_trust

(I) Explanation_Type	(J) Explanation_Type	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	.154	.199	1.000	-.400	.708
	3	-.385	.251	.454	-1.082	.313
2	1	-.154	.199	1.000	-.708	.400
	3	-.538*	.129	.004	-.898	-.179
3	1	.385	.251	.454	-.313	1.082
	2	.538*	.129	.004	.179	.898

Based on estimated marginal means

Figure 7: Results of the quantitative analysis

3 Awareness of Data Status

Analysis of population dynamics and identification of trends require users to be aware of the underlying processes, uncertainty and potential errors (Section 2.2). However, including data provenance in the Fish4Knowldege tool influences the usability and could complicate the information processing. Now marine biologists have to pay attention to the possible sources of errors, learn new technical concepts and incorporate them in their research. Data presented in the interface has to be transformed into information useful to biologists. To do so, they must perceive information, interpret and make conclusions on the basis of the perceived and transformed information [23]. Endsley indicates the possibility of an *information gap* when only part of the information could be processed by users. This cognitive process could cause errors in understanding and, eventually, lead to wrong conclusions.

Information processing is strongly related to the concept of *situation awareness*, "knowing what is happening around you" [9] or the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future [5]. Good situation awareness is necessary for effective human performance and decision making. Problems with it are frequently caused by data overload, non-integrated data and complex systems that are poorly understood. The design of complex systems should facilitate the situation awareness of important aspects of a situation.

Endsley distinguishes 3 levels in situation awareness related to information processing [6]:

- **Perception** of cues is fundamental. Without basic perception of fundamental information users are more likely to form an incorrect picture of the situation.
- **Comprehension** is an integration of multiple pieces of information and a determination of their relevance to the user's goals. Situation awareness encompasses how users combine, interpret, store and retain information.
- **Projection** is the ability to forecast future situation events and dynamics. This level of situation awareness is particular to highly skilled experts.

Situation awareness benefits from correctly prioritized information in the interface and timely directed attention. Jones [13] showed situations where all the necessary information was present, but a user did not pay attention to it. Human attention plays an important role in selecting the information on the screen. Several factors contribute to attention: 1) salient features; 2) expectancy to find the element; 3) value - importance of the events received at different locations and 4) effort - required to move attention in the environment [23]. The non-observance of these requirements can lead to *attention narrowing* and errors (e.g. a user may not notice that the filter value has been changed and continue analysis on a wrong dataset). In turn, attention can be influenced by both the perceptual salience of environmental cues and the meaningful direction of user's attention [10, 12].

Humans are limited not only by their attention, but also by working memory [23]. Wickens indicates that memory-loss (where information was initially perceived and then forgotten) could happen if there are delays between receiving information and using it, e.g. due to distractions or confusing layout. Working memory becomes of crucial importance in novel systems such as Fish4Knowledge when users are not familiar with the used technique and must combine, interpret and apply unusual information within the limits of their working memory.

A user's mental model also contributes to directing attention in efficient ways and providing meaningful information without loading working memory [9]. Mental models develop with experience and form expectations about the behaviour of a specific tool. Marine biologists do not have experience with computer vision software and their knowledge and expectations about marine biology could direct their attention and influence their information perception [22]. Expectations could

also be formed based on the prior experience, instructions and tutorials [9].
Human information processing is seen as alternating between data-driven and goal driven processing, which is critical for situation awareness [5, 8]. In the Fish4Knowledge interface, biologist direct their attention in accordance with their goals (goal driven processing) and actively seek needed information. They should be able to react on perceived environmental cues indicating that they need to activate other goals (e.g., check whether the number of samples was enough in the analysed dataset or the software performs well for the chosen species).
In our study we assess the performance of the Fish4Knowledge interface in facilitating the situation awareness focusing on meaningful direction of *attention*, effective usage of *working memory*, fulfilling biologists' *expectations* and activating *goals* when necessary.

3.1 Method

3.1.1 Task development

We adapted the situation awareness global assessment technique (SAGAT) [6, 11] to obtain measurements of situation awareness. Based on the user requirements collected by the CWI team [2] and our user interviews from the study of data provenance needs (Section 2), we chose one of the tasks supported by the Fish4Knowledge tool: analysing patterns of fish occurrence from different species. Following the *Goal-Directed Task Analysis* format [7], we constructed the model of this task based on the scenarios and expert knowledge in video analysis results evaluation. While user observations play an important role in goal-directed task analysis, they are not applicable to our domain, since Fish4Knowledge users do not have experience in using similar software. We assume that the developed model could change once marine biologists start using the software in their current practices.
According to the task model, to take effective decisions users require good understanding of the ongoing processes. We decomposed their situation awareness for each decision into 3 questions, which correlate to the Endsley's levels of situation awareness: (*Perception*, *Comprehension* and *Projection*). For example, the list below shows that to take the decision *Which period of year has the most fish?* users require the following information:

- SUBGOAL: Decide whether the observed pattern depends on time of year.
  1. *DECISION: Which period of year has the most fish?*
     (a) In which period of the year can we observe the most abundant fish population? (*Projection*)
     (b) What period of year has the most fish? (*Comprehension*)
     (c) What is the number of fish occurring during the specific week of the year? (*Perception*)

The complete version of the task model is given in Appendices B and D.

3.1.2 Setup

We developed a high-fidelity prototype of the Visualization tab to evaluate user performance for the chosen scenario (Figure 8). In this version of the prototype a user could visualize *Fish Counts*, *Normalized Fish Counts* and *Video Counts* on the y-axis; *Week of Year* and *Hour of Day* on the x-axis. The available visual representations are a line chart (Fig. 8) and a stacked chart. In the form of a stacked chart a user could visualize the composition of species over *Hours of Day* or *Weeks of Year*.
We distributed tasks over 3 *Situations* (Table 3.1.2). *Situation* is a modification of a *freezing probe* [6]: it simulates the moment in the working process of a researcher. It is a state of the interface

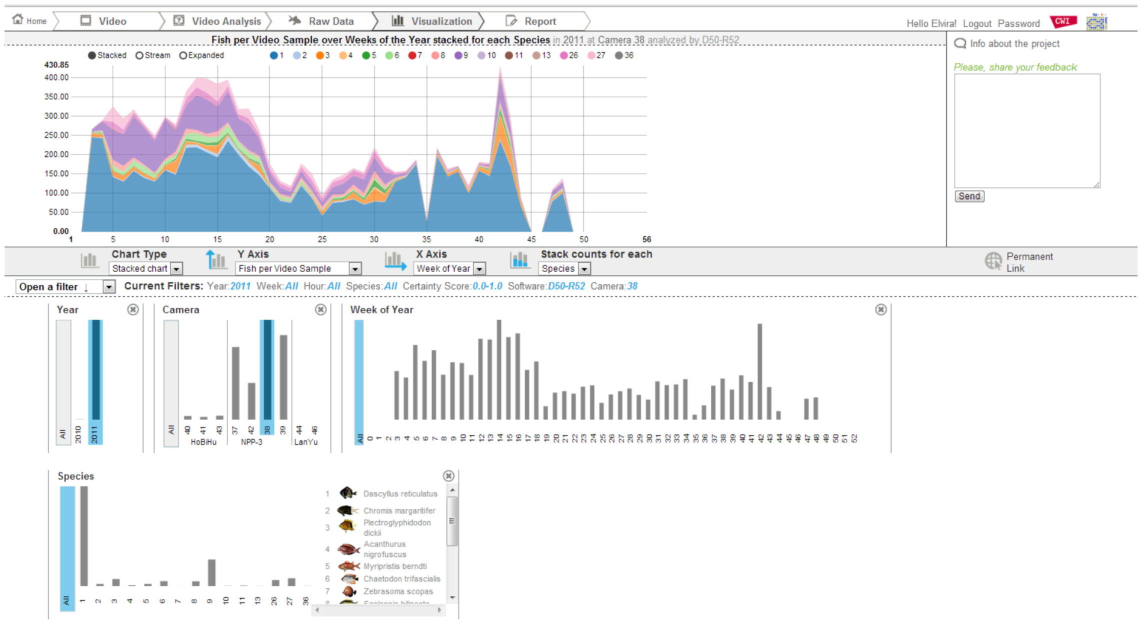


Figure 8: The Visualization tab showing the distribution of fish counts over weeks of the year for camera 38 in 2011

with predefined filter settings, specific values on x- and y-axes, opened filter widgets and chosen type of visualization. The *Situations* differ in the following criteria:

Levels of situation awareness	Situation 1	Situation 2	Situation 3
Perception	+		
Comprehension	+	+	
Projection	+	+	+

Table 3: Distribution of levels of situation awareness over the Situations

- the number of applied filters;
- the relevance of the applied filters to the questions asked;
- the relevance of the opened filtering widgets to the applied filters;
- the amount of manipulations that users have to make to get the answer;
- the levels of situation awareness (e.g. *Situation 3* contains only questions from the Projection level).

The full list of questions and the description of the *Situations* is available in Appendices C and D.

3.1.3 Participants

We recruited 12 participants from Taiwanese Marine Biology institutions (5 professors, 5 researchers, 1 PhD candidate and 5 Master's students). We chose users whose field of research is related to coral reef fishery and who potentially could use the Fish4Knowledge system. The participants belong to one of the following groups:

- *Coral reef fishery* Biologists from three teams, Academia Sinica (Taiwan), National Museum of Marine Science and Technology (Taiwan) and Wageningen University (The Netherlands), study population dynamics and fish systematics of coral reef fish. They traditionally use diving observations to collect data. Some teams already use video cameras to collect data. For example, one team uses baited stereoscopic cameras to obtain samples from different locations. Another team dives with lighted hand-held cameras and moves along the predefined paths. The teams manually analyse the videos and would potentially be interested in applying video analysis methods to avoid manual counting. We presented F4K to the biologists whose research is focused on fish systematics, morphology and evolution

- *Aquatic toxicology* We interviewed a biologist based in the National Museum of Marine Biology and Aquarium (NMMBA) whose research is focused on ecotoxicology, animal behavior and biodiversity. They study the effects of manufactured chemicals and other anthropogenic and natural materials and activities on aquatic organisms at various levels of organization

- *Marine biology education* The team of biologists based in NMBBA monitor fish population and species composition in the aquarium to maintain the healthy balance of the artificial marine ecosystem. This aquarium is also a platform where the marine researchers conduct studies. The team also creates educative interactive programs to bring awareness to the Marine Aquarium visitors.

- *Phytoplankton, Microorganisms* Phytoplankton is a base of all food chains and a key component of coral reef ecosystems [6].

3.1.4 Procedure

In the beginning of the study we conducted a semi-structured interview focusing on possible research questions that the Fish4Knowledge tool could help to investigate. We focused on the techniques to handle uncertainty that are currently used in the marine biology domain.
Prior to the experiment we provided a tutorial that lasted 15-20 minutes on how to use the tool and let the users interact with the software to familiarize themselves with it. We provided an overview of possible visualizations and manipulations with the user interface (e.g. changing axes, filtering data). In our explanation we highlighted the following items:

- *Number of Video Samples, Fish Counts Normalization.* Number of processed videos in the dataset could vary due to different technical issues (e.g. encoding problems, videos to be processed, damaged camera). Users have an option to use fish counts either absolute or normalized by the number of analysed videos.

- *Certainty Scores* We explained the concept of certainty scores and manipulations with the related filtering widget in the interface.

- *Manipulations with Axes and Visualization Types* We provided information on how to obtain different types of visual representations and subsets of data.

- *Filters Summary* We showed the types of filter widgets available for the dataset and explicitly indicated where to see the overview of the applied filters.

After the tutorial we allowed users to interact with the interface for 3-5 minutes. We asked users to accomplish a set of tasks using the interface prepared for them. We consequently exposed them to the 3 different *Situations* (Table 3.1.2). After the third set of tasks we conducted a short semi-structured interview focusing on their experience with the tool.

3.1.5 Analysis

Users interactions and remarks we documented by two interviewers. We synthesized them into a list of issues related to different parts of the interface (e.g. *interactions with widgets* and *manipulations with axes*).

We compared users responses in the questionnaire to the expert answers and calculated the percentage of correct ones for each level of situation awareness.

The incorrect answers and the list of issues mentioned above were coded according to the aspects of situation awareness:

- *Goal Achievement*. We included all the cases where the high-level goal had not been achieved, due to inactivated sub-goals.
- *Attention*. We included problems with attention tunnelling and misplaced salience. For example, a user did not see the information that had been on the screen. The cues that the system provided to activate sub-goals were not understood by a user.
- *Working Memory* We included problems when users had the information, but did not activate it.
- *Expectations* We included cases when provided interface responses and information architecture did not fit users expectations about the tool.

3.2 Results

Users performed worst at the *Projection* level of situation awareness (Fig. 3.2). The majority of the errors on this level were caused by a not activated sub-goal to check the validity of the observed pattern. The errors on the *Perception* and *Comprehension* levels could mostly be explained by *attention problems* (the users did not see the information on the screen), *memory-loss* (they forgot the information that was shown to them in the tutorial) or unexpected interface response.

Goal Achievement Almost none of the users activated the sub-goal of verifying the observed patterns. For instance, they replied that no fish had been observed during a specific week of the year without checking the number of processed videos. That specific week our algorithms did not analyse any video samples, which simply meant that there was no data, but possibly could have been fish.

The software did not provide sufficient cues to check whether the dataset has even number of videos. We explained the difference between "Fish Counts" and "Fish per Video" and all participants agreed that Fish Count should not be used in most of the scenarios. However, only one biologist in one question changed the y-axis to "Fish per Video", whereas the other participants did not consider the normalization.

In the question about whether the number of observed video samples were enough for scientific analysis only 1 participant checked the quality of the samples by opening the certainty scores widget.

Attention The majority of attention problems were caused by focusing it on the main visualization or filtering widgets (*tunnelling*). For example, in *Situation 3*, none of the users considered

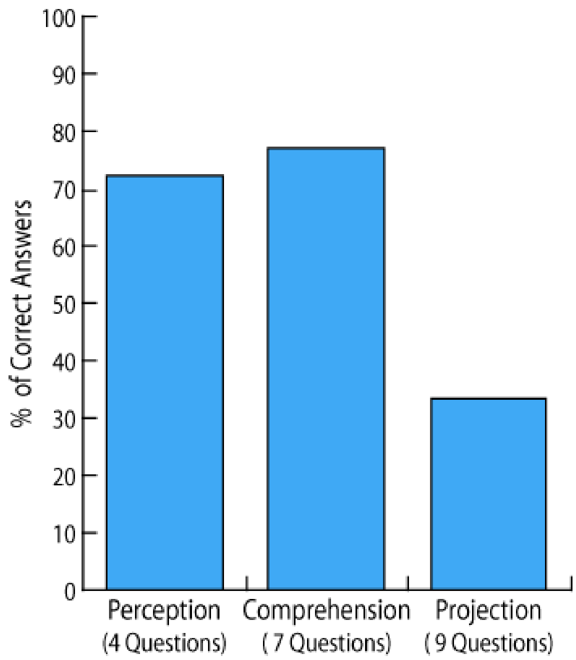


Figure 9: Correct answers given by the participants on different levels of situation awareness

that the data was collected only for 11 a.m. and incorrectly interpreted the data. Only when the visualization looked abnormal (Fig. 3.2), they explored what subset of the data they were currently analysing.
Interacting with widgets revealed one more attention problems. The users expected that by clicking on a specific value in the filter widget would automatically cancel their previous choice, whereas the system just filtered the data on two parameters (e.g. when week 13 is already selected, clicking on week 12 in the filtering widget would filter the data on week 12 & 13). Some participants did not notice 2 parameters instead of 1 in the *filters summary* zone and continued analysis with an incorrect dataset.

A number of the problems that could also be related to attention were caused by poor usability of the visualization. The graduations on the x-axis and y-axis marked every fifth division (e.g. 0, 5, 10) and users had to use a pen or a finger to count the specific week or day. In case if they did not use any extra object to draw a projection on y-axis they made a mistake (e.g. 250 fish instead of 270). Also, changing a scale on the y-axis was not noticed by some of the users. It caused the false perception of [almost] constant number of fish regardless of the applied filters. One user did not notice the "Fish per Video" y-axis menu option and calculated the average number of fish by using the other two menu options: "Fish Count" and "Video Count".

Working memory & Expectations Users did not have 1 specific scenario in mind while answering the questions, which could be more appropriate for studying the working memory issues. They did not have a well-developed mental model of the software behaviour and had never interacted with similar computer vision based software. In our analysis we joined these two types of problems, because it is difficult to determine and separate the real cause: whether users forgot where to find the information or have different expectations based on the provided cues. Thus, half of the participants were confused with finding the number of analysed video samples, which caused

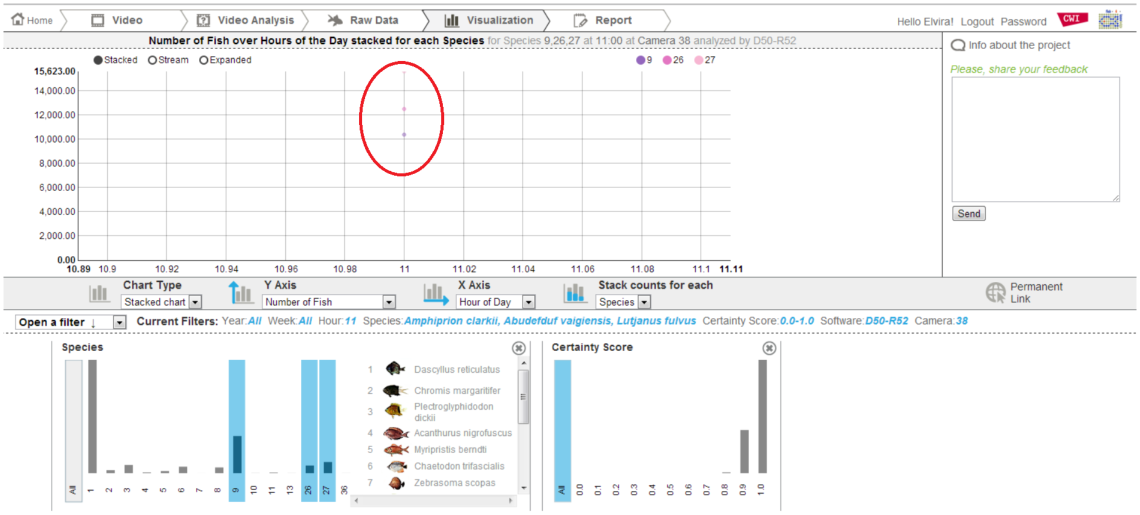


Figure 10: Attention tunnelling caused confusion in data interpretation

mistakes in their responses. They tried to find the information in the *Video* tab. In the absence of the explicit answer, they started counting the videos that were analysed for each camera.
The axes of the graph were not labelled, since their names were indicated in the dropdown menus below it (Fig. 8). The users complained that they *forget what they chose in the dropdown menus before*. We considered it as both *working memory* issue - when the time between receiving and using information was too long and unsatisfied *expectations*, because the users did not expect to see them at that place.
Overall, the users performed well at levels 1(72,5%) and 2 (77 %) of situation awareness. This indicates that the software potentially could provide good *situation awareness* for the users. The performance on Level 3 was much worse than at the other levels (33,2 %). None of the participants mastered the concept of verifying the observed fish counts by checking the samples' quality and amount. According to Endsley [9], the ability to forecast future dynamics and understand all the underlying mechanisms is a mark of a skilled expert. None of the participating marine biologists could be considered as skilled in operating computer vision based software yet.

3.3 Limitations

The users did not have a specific goal (Section 3.1) in mind while answering the questions and, probably, did not have a correct mental model. This could explain that almost none of the users checked the validity of the correlation unless it was explicitly asked.
In some cases we observed the misunderstanding of the questions due to 1) language issues; 2) the terminology in the questionnaire was not the same that they are used to.
Provided tutorial is not a replacement of a long term experience.

3.4 Discussion & Conclusion

Uncertainty in data influences user information processing, as they need to consider sample size and errors in the analysed dataset. They risk becoming overwhelmed by the complexity of the underlying processes and overloaded by information that has to be considered when taking decisions (e.g., applied filters, software evaluation, samples evenness and quantity). Our case study revealed that users do not tend to check the validity of the patterns that they observe, which leads them to invalid conclusions. To overcome this problem, the indications of the uncertain or erroneous data could be provided on the main visualization along with the information about the sample size. For example, instead of switching between two different y-axes (*Fish Counts* and *Number of Video Samples*), both lines could be displayed on one visualization or the line of fish counts could be color-coded to show the approximate number of the samples used. Besides improving the awareness of the uncertainties in the data, it will guide people's attention to the most "questionable" parts of the data.

Apart from data provenance, dataset filters also contributed to wrong conclusions made by users. Making information about the applied filters more salient in the Visualization tab could help users to be aware of the dataset status. For example, animating the parameters change, including the parameters in the title of the visualization.

The participants showed good performance in finding the information on the screen and comparing the results of several measurements. Their performance in forecasting the future dynamics, would probably improve with experience. Bad performance, especially, working memory issues, could have been impacted by length or quality of the provided tutorial. According to Endsley [9], automaticity could improve their performance on all levels, as they will be familiar with concepts that they operate with.

4 Design Implications for Tools Based on Automated Video Analysis

Following the findings from the case studies discussed in Sections 2, 3, we derived a number of design implications for tools based on video analysis. Provided list of guidelines is relevant for the users who need to draw *scientifically valid* conclusions from uncertain data and who does not have a technical background.

- **Information Architecture**
  - *Provide data provenance information.* Researchers need information about uncertainty inherent to the dataset (Section2.2). This should include potential errors, a description of the sampling method, camera location and field of view. The description of potential errors must support users in differentiating trends from noise in the data, and identifying meaningful patterns. For the marine biology community, it is essential to provide well-accepted measurements of uncertainty, such as means and standard deviations over samples, which can be presented in the form of error bars.
  - *Ground truth collection should be accessible for users* (Section 2.2, [2]). Ground truth images allow users to control the differences with the pictures in the analyzed dataset (e.g., changes in the field of view, fish abundance, water/image quality, differences in species composition compared to the overall collection). Marine biologists are aware of the errors that expert users, manually labelling ground truth images, could introduce. Browsing ground truth collection helps to control the coverage of the video collection and of the extracted features easier.
  - *Provide verification of the algorithms by fellow biologists* (Section 2.2). The majority of the interviewed marine biologists were not experts in technical aspects of video analysis. They expressed a need for verification of the algorithms by fellow-biologists with expertise in statistics.

- *Provide performance of the software in different conditions* (Section 3) Decompose the evaluation of the software into meaningful conditions for the domain. For example, in marine biology domain, decomposition could be over water conditions (e.g., murky, covered by algae), time of the day (e.g., morning with bright sunlight or evening with lack of light) or background (e.g. open sea, rocks, corals). More specific evaluation could help users to reduce their doubts about the uncertainty and know more precisely what could bias *their* dataset (e.g., in murky water the software could have worse performance and therefore biologists would know that winter time measurements could be biased).

- **Information Processing**
  - *Highlight parameters of a dataset* (Section 3). Due to the memory-loss, attention tunnelling, users tend to forget the filters that they have applied while working with the dataset. It could lead to wrong conclusions, since the users could overlook what parameters are applied. Users attention needs to be drawn to the parameters of the filters in the Visualization tab.
  - *Provide explicit cues to verify the validity of the correlation in the Visualization tab* (Section 3). Users do not check the validity of the pattern while working with the visualization. We recommend explicitly indicating cues (e.g., mark parts of the visualization where no videos were analysed in the main graph). While indicating cues on the visualization could potentially help to improve awareness of uncertainties in the dataset, further research has to be done to investigate the effects of such advanced visualization on human information processing.

5 Conclusions

We highlight challenges in designing video analysis tools for scientific use in the framework of the Fish4Knowledge project. Uncertainty and errors introduced by the software influence user trust in the video analysis technique. One way to overcome trust issues is to make data provenance information accessible to users. Revealing technical details of software evaluation could affect user trust and understanding of the underlying processes. Our studies indicate that user trust in the technical competence of the software could potentially be improved by showing more extensive evaluation of the software. Marine biologists have difficulties understanding the concepts, but would still prefer to have more extensive explanations to short ones. Moreover, they require additional information on data provenance, such as camera settings, software performance in different conditions and sampling method. However, users do not feel the need to check the validity of observed fish counts while interacting with the software. The parameters of the dataset could also be overlooked due to attention tunnelling. Despite the low understanding and trust, participants showed high acceptance of the tool.

The presented results could be relevant for scientifically-oriented video analysis tools. Future research could be done in exploring data visualization techniques that direct users attention to data provenance and visualization of potential biases. This study could contribute to exploration of problems related to over-/ under-confidence in the decisions made using the Fish4Knowlege tool.

References

[1] Donovan Artz and Yolanda Gil. A survey of trust in computer science and the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):58–71, 2007.

[2] E Beauxis-Aussalet, L Hardman, and J van Ossenbruggen. Fish4knowledge deliverable d2.1 user information needs.

[3] E Beauxis-Aussalet, J He, C Spampinato, B Boom, J van Ossenbruggen, and L Hardman. Fish4knowledge deliverable d2.3 component-based prototypes and evaluation criteria.

[4] Jeremy J Carroll, Christian Bizer, Pat Hayes, and Patrick Stickler. Named graphs, provenance and trust. In *Proceedings of the 14th international conference on World Wide Web*, pages 613–622. ACM, 2005.

[5] Mica R Endsley. Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 32, pages 97–101. SAGE Publications, 1988.

[6] Mica R Endsley. Situation awareness global assessment technique (sagat). In *Aerospace and Electronics Conference, 1988. NAECON 1988., Proceedings of the IEEE 1988 National*, pages 789–795. IEEE, 1988.

[7] Mica R Endsley. Direct measurement of situation awareness in simulations of dynamic systems: Validity and use of sagat. *Experimental analysis and measurement of situation awareness*, pages 107–113, 1995.

[8] Mica R Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1):32–64, 1995.

[9] Mica R Endsley. Theoretical underpinnings of situation awareness: A critical review. *Situation awareness analysis and measurement*, pages 3–32, 2000.

[10] Mica R Endsley and Mark D Rodgers. Distribution of attention, situation awareness, and workload in a passive air traffic control task: Implications for operational errors and automation. Technical report, DTIC Document, 1997.

[11] Mica R Endsley, Stephen J Selcon, Thomas D Hardiman, and Darryl G Croft. A comparative analysis of sagat and sart for evaluations of situation awareness. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 42, pages 82–86. SAGE Publications, 1998.

[12] Mica R Endsley and Robert P Smith. Attention distribution and decision making in tactical air combat. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 38(2):232–249, 1996.

[13] Debra G Jones and Mica R Endsley. Sources of situation awareness errors in aviation. *Aviation, Space, and Environmental Medicine*, 1996.

[14] Kari Kelton, Kenneth R Fleischmann, and William A Wallace. Trust in digital information. *Journal of the American Society for Information Science and Technology*, 59(3):363–374, 2008.

[15] Bertram Ludäscher, Ilkay Altintas, Chad Berkley, Dan Higgins, Efrat Jaeger, Matthew Jones, Edward A Lee, Jing Tao, and Yang Zhao. Scientific workflow management and the kepler system. *Concurrency and Computation: Practice and Experience*, 18(10):1039–1065, 2006.

[16] Clifford A Lynch. When documents deceive: trust and provenance as new factors for information retrieval in a tangled web. *JASIST*, 52(1):12–17, 2001.

[17] Maria Madsen and Shirley Gregor. Measuring human-computer trust. In *Proceedings of Eleventh Australasian Conference on Information Systems*, pages 6–8. Citeseer, 2000.

[18] Carolyn Mair, Gada Kadoda, Martin Lefley, Keith Phalp, Chris Schofield, Martin Shepperd, and Steve Webster. An investigation of machine learning based prediction systems. *Journal of Systems and Software*, 53(1):23–29, 2000.

[19] Daniel J McAllister. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal*, 38(1):24–59, 1995.

[20] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R Pocock, Anil Wipat, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–3054, 2004.

[21] Saeedeh Shekarpour and SD Katebi. Modeling and evaluation of trust with an extension in semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(1):26–36, 2010.

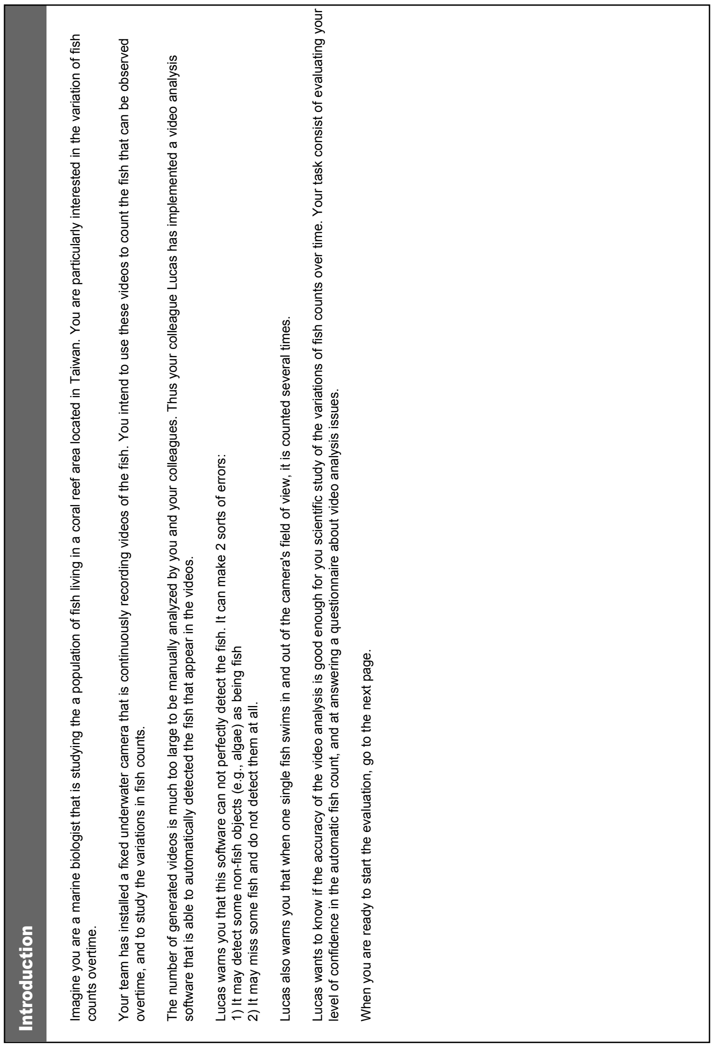
[22] C.D. Wickens. *Engineering Psychology and Human Performance.* New York:Harper Collins, 1992.

[23] Christopher D Wickens and C Melody Carswell. Information processing. *Handbook of human factors and ergonomics*, 2:89–122, 1997.

[24] L Richard Ye and Paul E Johnson. The impact of explanation facilities on user acceptance of expert systems advice. *Mis Quarterly*, pages 157–172, 1995.

Appendices

Appendix A Tasks Used to Measure User Trust and Understanding



Introduction

Imagine you are a marine biologist that is studying the a population of fish living in a coral reef area located in Taiwan. You are particularly interested in the variation of fish counts overtime.

Your team has installed a fixed underwater camera that is continuously recording videos of the fish. You intend to use these videos to count the fish that can be observed overtime, and to study the variations in fish counts.

The number of generated videos is much too large to be manually analyzed by you and your colleagues. Thus your colleague Lucas has implemented a video analysis software that is able to automatically detected the fish that appear in the videos.

Lucas warns you that this software can not perfectly detect the fish. It can make 2 sorts of errors:
1) It may detect some non-fish objects (e.g., algae) as being fish
2) It may miss some fish and do not detect them at all.

Lucas also warns you that when one single fish swims in and out of the camera's field of view, it is counted several times.

Lucas wants to know if the accuracy of the video analysis is good enough for you scientific study of the variations of fish counts over time. Your task consist of evaluating your level of confidence in the automatic fish count, and at answering a questionnaire about video analysis issues.

When you are ready to start the evaluation, go to the next page.

Task 1/4 - Evaluating the video analysis software

Lucas has made a presentation of his software, and an evaluation the quality of the automatic fish count. He wants you to read an online presentation of his results.

To do so, please follow this link: [URL]

When you are done with reading the online presentation, please go to the next page.

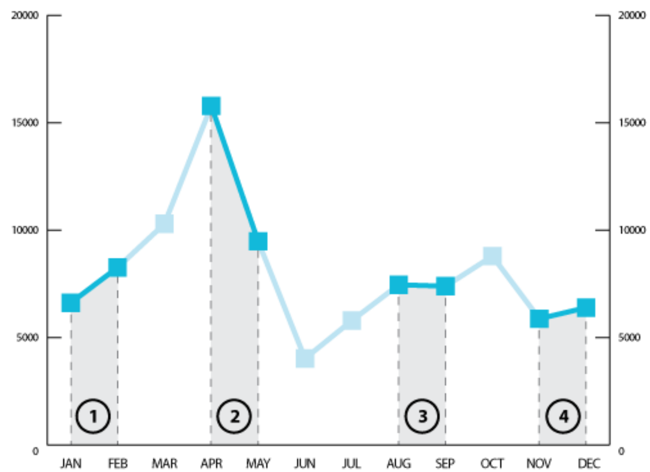
Task 1/4 - Questionnaire p.1/4

***1. What is your level of confidence in the video analysis software?**
Please indicate how much you agree with the following statements.

	I strongly disagree	I disagree	I tend to disagree	I am neutral	I tend to agree	I agree	I strongly agree
The video analysis software produces accurate counts of fish.	○	○	○	○	○	○	○
The automatic fish counts are trustworthy.	○	○	○	○	○	○	○
The accuracy of the video analysis software is good enough to be used for the scientific study of fish counts over time.	○	○	○	○	○	○	○
The video analysis software uses an appropriate method for analyzing the videos and counting fish.	○	○	○	○	○	○	○
The automatic fish counts produced by the video analysis software are as good as the fish counts that marine biology experts could produce.	○	○	○	○	○	○	○
The system correctly handles the errors it produces.	○	○	○	○	○	○	○
To have more trust in the video analysis software, I would need more explanations about how the software works.	○	○	○	○	○	○	○

Do you have any other question? or any additional information you would need?

We focus on the four following trends that can be observed in the automatic fish counts for 2011.



***2. What is your level of confidence in the above trends?**

The trends observed in the automatic count can be different than the trends occurring for real in the ecosystem. This is particularly due to the errors introduced by the video analysis software.

For each trend identified above, please indicate how you would qualify the intensity of the trend (e.g., small or important increase. And please also indicate how confident you are in the fact that the observed trend is exactly the same in reality.

	What trend is it?	How likely the trend is the same in reality?
Trend 1 - Jan. - Feb.		
Trend 2 - Apr. - May		
Trend 3 - Aug. - Sept.		
Trend 4 - Nov. - Dec.		

Do you have any remark about the difference between observed trends and the trends in reality?

Task 1/4 - Questionnaire p.2/4

*3. Understanding the system. Please indicate how much you agree with the following statements.

	I strongly disagree	I disagree	I tend to disagree	I am neutral	I tend to agree	I agree	I strongly agree
I fully understand how the video analysis software works.	○	○	○	○	○	○	○
It is easy to understand how the video analysis software works.	○	○	○	○	○	○	○
I want more information about how the video analysis software works, about its inner mechanisms, and about the techniques involved in the automatic fish detection.	○	○	○	○	○	○	○
I understand what errors are produced by the video analysis software.	○	○	○	○	○	○	○
I understand how the errors produced by the video analysis software can influence the results of my scientific study of fish counts.	○	○	○	○	○	○	○
I understand how to handles these errors and minimize their influence on my scientific research.	○	○	○	○	○	○	○

Do you have any other question? or any additional information you would need?

*4. Quality of the explanations. Please indicate how much you agree with the following statements.

	I strongly disagree	I disagree	I tend to disagree	I am neutral	I tend to agree	I agree	I strongly agree
I fully understood the explanations given about the video analysis software.	○	○	○	○	○	○	○
The explanations given about the video analysis software were easy to understand.	○	○	○	○	○	○	○
I was interested in the explanations given about how the video analysis software works.	○	○	○	○	○	○	○
The given explanations contained enough information for understanding how the video analysis software works.	○	○	○	○	○	○	○
I want more information about how the video analysis software works.	○	○	○	○	○	○	○

Additional remarks are welcome:

Task 1/4 - Questionnaire p.3/4

To improve the accuracy of the video analysis software, Lucas has developed and tested several versions of the software. The different versions can produce different counts of fish. Lucas has chosen the version that produces the fish count that is the closest to the fish count produced by experts.

***5. Amongst the versions of the video analysis software, do you that Lucas can obtain the following results:**

	Yes, it is possible	No, it is impossible	N/A
The automatic fish counts can be superior to the fish counts manually produced by experts (i.e., Automatic Count > Expert Count).	○	○	○
The automatic fish counts can be inferior to the fish counts manually produced by experts (i.e., Automatic Count < Expert Count).	○	○	○

***6. Why automatic counts can be different from expert count?**
Amongst the following statements, please indicate which are true. Please also indicate which statements can possibly explain why the automatic count may not be the same as the manual count.

	Is this true?	Can it cause the automatic count to be different from the expert count?
The automatic fish counts is likely to contain non-fish objects (e.g., rocks) that are incorrectly considered as being a fish.		
When one single fish swims in and out of the camera's field of view, it is counted several times by the video analysis software. It is as well counted several times in the manual fish count.		
The automatic fish counts is likely to miss some fish that are not detected at all.		

Do you think of any other reason that could explain why the automatic counts are different from the experts' count?

***7. Is there other sources of errors?**

The video analysis errors were evaluated for a small set of Videos for Evaluation (as shown in the "Video" tab). The video analysis software is applied to a much larger set of videos to process. When analyzing a the larges set of video, do you think that it is possible to encounter the following sources of errors? And would you like to measure and verify these errors?

	Can we encouter these errors?	Do you want to evaluate the importance of these errors?
Some video may be missing due to errors during the recording of the video.		
Some video may be of very poor quality due to video encoding errors.		
Some video may be of very poor quality due to the presence of dirt or algae on the camera lens.		
Some videos may not be analyzed at all due to video processing errors.		
Some camera's filed of view may have changed (e.g., due to strong current).		
Non-fish objects may appear more often in the larger set of videos.		

Do you think of any other source of error?

Task 1/4 - Questionnaire p.4/4

***8. The image below the counts of fish obtained by 2 different versions of the video analysis software that Lucas has developed. Which is the most accurate version of the software?**

	Version A	Version B	I don't know
	○	○	○

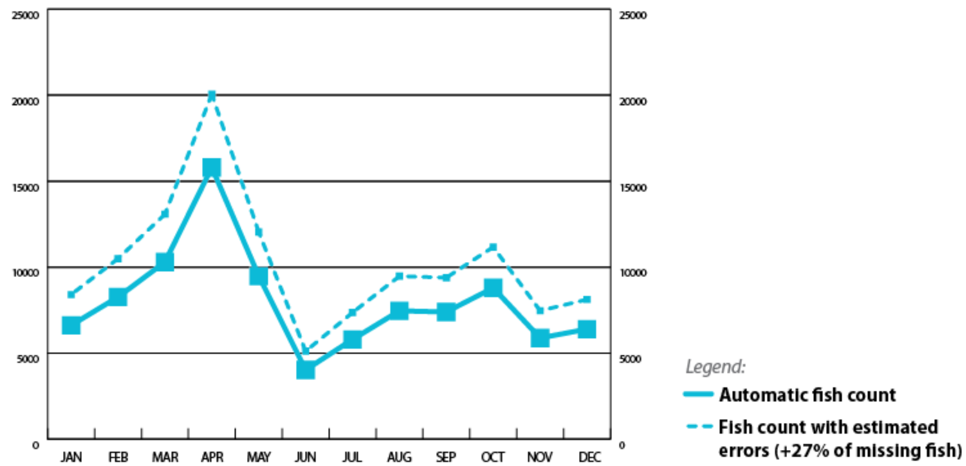
Why did you choose this version?

***9. The image below the counts of fish obtained by 2 different versions of the video analysis software that Lucas has developed. Which is the most accurate version of the software?**

Version A	Version C	I don't know
○	○	○

Why did you choose this version?

The image below show the fish counts for the year 2011, and a report of the errors made in the set of Video for Evaluation.



***10. Which fish count would you choose to use for studying the variation of fish count over time?**

- ○ I would choose the automatic count without a report of the errors (the solid blue line above).
- ○ I would choose the automatic count with report of the errors (the dashed line above).
- ○ I don't know.

Why would you choose these fish counts?

***11. Is there other source of errors?**

The evaluation of the video analysis errors is made for a small set of Videos for Evaluation (as shown in the "Video" tab). But the video analysis software is applied to a much larger set of videos collected for the year 2011. When analyzing a larger set of videos, do you think that it is possible to encounter the following sources of errors. And would you like to measure and verify these possible errors.

	Is it possible these errors	Can it cause the automatic count to be different from the expert count?
The automatic fish counts is likely to contain non-fish objects (e.g., rocks) that are incorrectly considered as being a fish.		
When one single fish swims in and out of the camera's field of view, it is counted several times by the video analysis software. It is as well counted several times in the manual fish count.		
The automatic fish counts is likely to miss some fish that are not detected at all.		

Do you think of any other reason that could explain why the automatic counts are different from the experts' count?

Task 2/4 - Explore the User Interface

You have completed the 1st task.

For your 2nd task, you will have to explore a new version of the presentation of the software results. The new presentation will give you more details about the errors contained in the automatic fish counts.

To access the presentation, please follow this link: [URL]

When you are done with reading the presentation, please go to the next page.

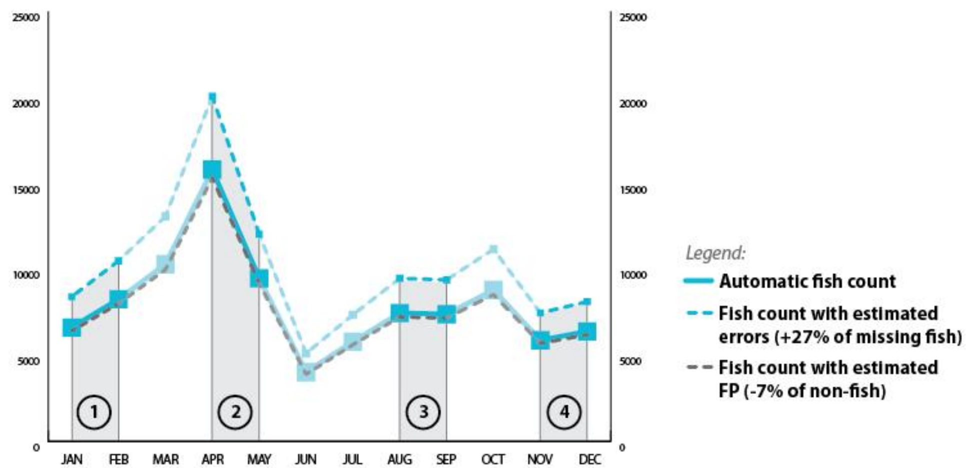
Task 2/4 - Questionnaire p.1/4

***12. What is your level of confidence in the video analysis software?**
Please indicate how much you agree with the following statements.

	I strongly disagree	I disagree	I tend to disagree	I am neutral	I tend to agree	I agree	I strongly agree
The video analysis software produces accurate counts of fish.	○	○	○	○	○	○	○
The automatic fish counts are trustworthy.	○	○	○	○	○	○	○
The accuracy of the video analysis software is good enough to be used for the scientific study of fish counts over time.	○	○	○	○	○	○	○
The video analysis software uses an appropriate method for analyzing the videos and counting fish.	○	○	○	○	○	○	○
The automatic fish counts produced by the video analysis software are as good as the fish counts that marine biology experts could produce.	○	○	○	○	○	○	○
The system correctly handles the errors it produces.	○	○	○	○	○	○	○
To have more trust in the video analysis software, I would need more explanations about how the software works.	○	○	○	○	○	○	○

Do you have any other question? or any additional information you would need?

We focus on the four following trends that can be observed in the automatic fish counts for 2011.



***13. What is your level of confidence in the above trends?**

The trends observed in the automatic count can be different than the trends occurring for real in the ecosystem. This is particularly due to the errors introduced by the video analysis software.

For each trend identified above, please indicate how you would qualify the intensity of the trend (e.g., small or important increase. And please also indicate how confident you are in the fact that the observed trend is exactly the same in reality.

	What trend is it?	How likely the trend is the same in reality?
Trend 1 - Jan. - Feb.		
Trend 2 - Apr. - May		
Trend 3 - Aug. - Sept.		
Trend 4 - Nov. - Dec.		

Do you have any remark about the difference between observed trends and the trends in reality?

Task 2/4 - Questionnaire p.2/4

*14. Understanding the system. Please indicate how much you agree with the following statements.

	I strongly disagree	I disagree	I tend to disagree	I am neutral	I tend to agree	I agree	I strongly agree
I fully understand how the video analysis software works.	○	○	○	○	○	○	○
It is easy to understand how the video analysis software works.	○	○	○	○	○	○	○
I want more information about how the video analysis software works, about its inner mechanisms, and about the techniques involved in the automatic fish detection.	○	○	○	○	○	○	○
I understand what errors are produced by the video analysis software.	○	○	○	○	○	○	○
I understand how the errors produced by the video analysis software can influence the results of my scientific study of fish counts.	○	○	○	○	○	○	○
I understand how to handles these errors and minimize their influence on my scientific research.	○	○	○	○	○	○	○

Do you have any other question? or any additional information you would need?

*15. Quality of the explanations. Please indicate how much you agree with the following statements.

	I strongly disagree	I disagree	I tend to disagree	I am neutral	I tend to agree	I agree	I strongly agree
I fully understood the explanations given about the video analysis software.	○	○	○	○	○	○	○
The explanations given about the video analysis software were easy to understand.	○	○	○	○	○	○	○
I was interested in the explanations given about how the video analysis software works.	○	○	○	○	○	○	○
The given explanations contained enough information for understanding how the video analysis software works.	○	○	○	○	○	○	○
I want more information about how the video analysis software works.	○	○	○	○	○	○	○

Additional remarks are welcome:

Task 2/4 - Questionnaire p.3/4

To improve the accuracy of the video analysis software, Lucas has developed and tested several versions of the software. The different versions can produce different counts of fish, and different numbers of False Positives, True Positives and False Negatives. Lucas has chosen the version of the software that produces the fish count that is the closest to the fish count produced by experts, while having the fewest False Positives and False Negatives.

*16. Please indicate if the following facts can explain the differences in numbers of False Positives, True Positives, and False Negatives that can be obtained for different versions of the software.

	This influences the number of False Positives	This influences the number of True Positives	This influences the number of False Negatives
Some versions of the software are more likely to detect non-fish objects (e.g., algae) as being a fish.	☐	☐	☐
Some versions of the software are more likely to correctly detect the fish in the videos.	☐	☐	☐
Some versions of the software are more likely to miss the detection of some fish in the videos.	☐	☐	☐

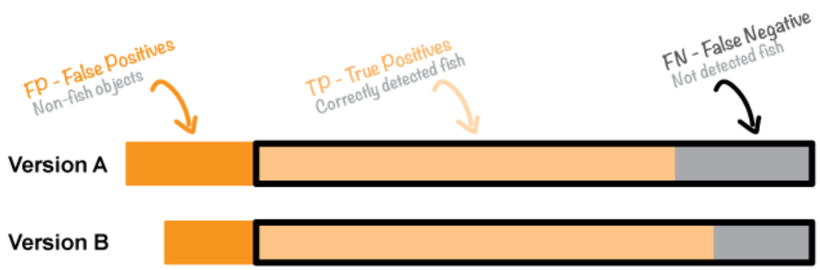
*17. Amongst the versions of the video analysis software, we compare the numbers of False Positives, True Positives, False Negatives. Do you that it is possible to find a version of the software that produces the following results.

	A > B	A < B	A = B
We compare A) the number of True Positives (TP); and B) the number of False Negatives (FN). Is it possible that:	☐	☐	☐
We compare A) the number of False Positives (FP); and B) the number of False Negatives (FN). Is it possible that:	☐	☐	☐
We compare A) the number of True Positives (TP); and B) the number of False Positives (FP). Is it possible that:	☐	☐	☐
We compare A) the manual fish count; and B) the sum of True Positives and False Negatives (TP+FN). Is it possible that:	☐	☐	☐
We compare A) the sum of True Positives and False Positves (TP+FP); and B) the sum of True Positives and False Negatives (TP+FN). Is it possible that:	☐	☐	☐

Task 2/4 - Questionnaire p.4/4

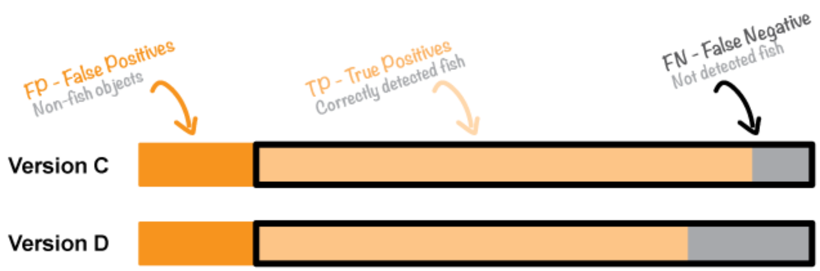
***18. The image below the counts of fish obtained by 2 different versions of the video analysis software that Lucas has developed. Which is the most accurate version of the software?**

Version A	Version B	I don't know
◯	◯	◯



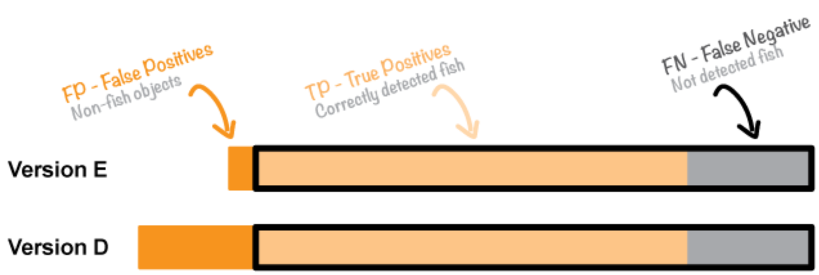
***19. The image below shows the counts of fish obtained by 2 different versions of the video analysis software that Lucas has developed. Which is the most accurate version of the software?**

Version C	Version D	I don't know
◯	◯	◯



*20. The image below the counts of fish obtained by 2 different versions of the video analysis software that Lucas has developed. Which is the most accurate version of the software?

Version E	Version D	I don't know
○	○	○



Task 3/4 - Explore the User Interface

You have completed the 2nd task.

For your 3rd task, you will now be given examples of the fish images that are used by our system. Your new task is the browse the examples of fish images of the fish that were automatically or manually detected.

You can access the examples of fish images at this URL: [URL]

When you are done with exploring the fish images, please go to the next page.

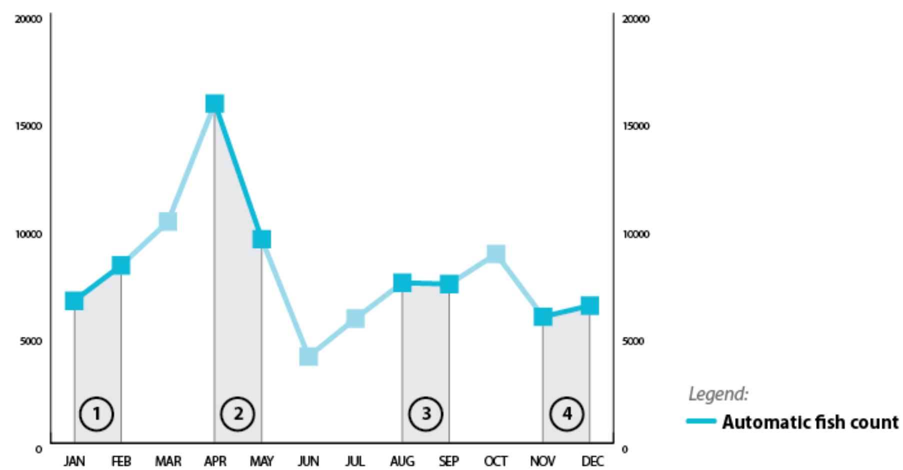
Task 3/4 - Questionnaire p.1/4

***21. What is your level of confidence in the video analysis software?**
Please indicate how much you agree with the following statements.

	I strongly disagree	I disagree	I tend to disagree	I am neutral	I tend to agree	I agree	I strongly agree
The video analysis software produces accurate counts of fish.	○	○	○	○	○	○	○
The automatic fish counts are trustworthy.	○	○	○	○	○	○	○
The accuracy of the video analysis software is good enough to be used for the scientific study of fish counts over time.	○	○	○	○	○	○	○
The video analysis software uses an appropriate method for analyzing the videos and counting fish.	○	○	○	○	○	○	○
The automatic fish counts produced by the video analysis software are as good as the fish counts that marine biology experts could produce.	○	○	○	○	○	○	○
The system correctly handles the errors it produces.	○	○	○	○	○	○	○
To have more trust in the video analysis software, I would need more explanations about how the software works.	○	○	○	○	○	○	○

Do you have any other question? or any additional information you would need?

We focus on the four following trends that can be observed in the automatic fish counts for 2011.



***22. What is your level of confidence in the above trends?**

The trends observed in the automatic count can be different than the trends occurring for real in the ecosystem. This is particularly due to the errors introduced by the video analysis software.

For each trend identified above, please indicate how you would qualify the intensity of the trend (e.g., small or important increase. And please also indicate how confident you are in the fact that the observed trend is exactly the same in reality.

	What trend is it?	How likely the trend is the same in reality?
Trend 1 - Jan. - Feb.		
Trend 2 - Apr. - May		
Trend 3 - Aug. - Sept.		
Trend 4 - Nov. - Dec.		

Do you have any remark about the difference between observed trends and the trends in reality?

Task 3/4 - Questionnaire p.2/4

*23. Understanding the system. Please indicate how much you agree with the following statements.

	I strongly disagree	I disagree	I tend to disagree	I am neutral	I tend to agree	I agree	I strongly agree
I fully understand how the video analysis software works.	○	○	○	○	○	○	○
It is easy to understand how the video analysis software works.	○	○	○	○	○	○	○
I want more information about how the video analysis software works, about its inner mechanisms, and about the techniques involved in the automatic fish detection.	○	○	○	○	○	○	○
I understand what errors are produced by the video analysis software.	○	○	○	○	○	○	○
I understand how the errors produced by the video analysis software can influence the results of my scientific study of fish counts.	○	○	○	○	○	○	○
I understand how to handles these errors and minimize their influence on my scientific research.	○	○	○	○	○	○	○

Do you have any other question? or any additional information you would need?

*24. Quality of the explanations. Please indicate how much you agree with the following statements.

	I strongly disagree	I disagree	I tend to disagree	I am neutral	I tend to agree	I agree	I strongly agree
I fully understood the explanations given about the video analysis software.	○	○	○	○	○	○	○
The explanations given about the video analysis software were easy to understand.	○	○	○	○	○	○	○
I was interested in the explanations given about how the video analysis software works.	○	○	○	○	○	○	○
The given explanations contained enough information for understanding how the video analysis software works.	○	○	○	○	○	○	○
I want more information about how the video analysis software works.	○	○	○	○	○	○	○

Additional remarks are welcome:

Task 4/4 - Explore the User Interface

You have completed the 3rd task.

For your last task, you will have to explore a new version of the presentation of the software. This new presentation will give you more details about the errors introduced by the video analysis software.

You can access the new presentation at this URL: [URL]

When you are done with reading the presentation, please go to the next page.

Task 4/4 - Questionnaire p.1/4

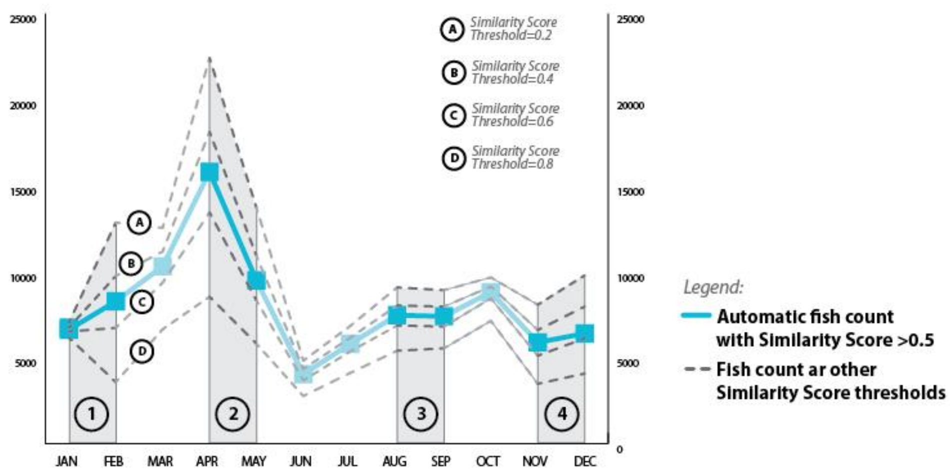
***25. What is your level of confidence in the video analysis software?**

Please indicate how much you agree with the following statements.

	I strongly disagree	I disagree	I tend to disagree	I am neutral	I tend to agree	I agree	I strongly agree
The video analysis software produces accurate counts of fish.	○	○	○	○	○	○	○
The automatic fish counts are trustworthy.	○	○	○	○	○	○	○
The accuracy of the video analysis software is good enough to be used for the scientific study of fish counts over time.	○	○	○	○	○	○	○
The video analysis software uses an appropriate method for analyzing the videos and counting fish.	○	○	○	○	○	○	○
The automatic fish counts produced by the video analysis software are as good as the fish counts that marine biology experts could produce.	○	○	○	○	○	○	○
The system correctly handles the errors it produces.	○	○	○	○	○	○	○
To have more trust in the video analysis software, I would need more explanations about how the software works.	○	○	○	○	○	○	○

Do you have any other question? or any additional information you would need?

We focus on the four following trends that can be observed in the automatic fish counts for 2011.



***26. What is your level of confidence in the above trends?**

The trends observed in the automatic count can be different than the trends occurring for real in the ecosystem. This is particularly due to the errors introduced by the video analysis software.

For each trend identified above, please indicate how you would qualify the intensity of the trend (e.g., small or important increase. And please also indicate how confident you are in the fact that the observed trend is exactly the same in reality.

	What trend is it?	How likely the trend is the same in reality?
Trend 1 - Jan. - Feb.		
Trend 2 - Apr. - May		
Trend 3 - Aug. - Sept.		
Trend 4 - Nov. - Dec.		

Do you have any remark about the difference between observed trends and the trends in reality?

Task 4/4 - Questionnaire p.2/4

*27. Understanding the system. Please indicate how much you agree with the following statements.

	I strongly disagree	I disagree	I tend to disagree	I am neutral	I tend to agree	I agree	I strongly agree
I fully understand how the video analysis software works.	○	○	○	○	○	○	○
It is easy to understand how the video analysis software works.	○	○	○	○	○	○	○
I want more information about how the video analysis software works, about its inner mechanisms, and about the techniques involved in the automatic fish detection.	○	○	○	○	○	○	○
I understand what errors are produced by the video analysis software.	○	○	○	○	○	○	○
I understand how the errors produced by the video analysis software can influence the results of my scientific study of fish counts.	○	○	○	○	○	○	○
I understand how to handles these errors and minimize their influence on my scientific research.	○	○	○	○	○	○	○

Do you have any other question? or any additional information you would need?

*28. Quality of the explanations. Please indicate how much you agree with the following statements.

	I strongly disagree	I disagree	I tend to disagree	I am neutral	I tend to agree	I agree	I strongly agree
I fully understood the explanations given about the video analysis software.	○	○	○	○	○	○	○
The explanations given about the video analysis software were easy to understand.	○	○	○	○	○	○	○
I was interested in the explanations given about how the video analysis software works.	○	○	○	○	○	○	○
The given explanations contained enough information for understanding how the video analysis software works.	○	○	○	○	○	○	○
I want more information about how the video analysis software works.	○	○	○	○	○	○	○

Additional remarks are welcome:

Task 3/4 - Questionnaire p.3/4

To improve the accuracy of the video analysis software, Lucas can choose an appropriate Similarity Score threshold. The different thresholds can produce different counts of fish, and different numbers of False Positives, True Positives and False Negatives.

Lucas has chosen the threshold the diminished the number of errors while keeping the automatic fish count as close as possible to the manual fish count.

*29. Please indicate if the following facts can explain the differences in numbers of False Positives, True Positives, and False Negatives that can be obtained for different Similarity Score thresholds.

	This influences the number of False Positives	This influences the number of True Positives	This influences the number of False Negatives
Some thresholds are more likely to discard non-fish objects (e.g., algae) that were detected as being fish.	☐	☐	☐
Some thresholds are more likely to include non-fish objects in the fish counts.	☐	☐	☐
Some thresholds are more likely to incorrectly discard fish that were correctly detected.	☐	☐	☐

*30. Amongst the Similarity Score threshold used, we compare the numbers of False Positives, True Positives, False Negatives. Do you that it is possible to find the following results.

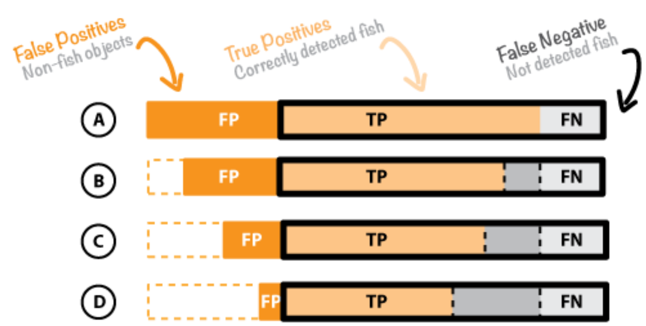
	A > B	A < B	A = B
We compare B) the number of True Positives (TP) for a threshold = 0.2; and B) the number of True Positives (TP) for a threshold = 0.6. Is it possible that:	☐	☐	☐
We compare B) the number of False Positives (FP) for a threshold = 0.2; and B) the number of False Positives (FP) for a threshold = 0.6. Is it possible that:	☐	☐	☐
We compare B) the number of False Negatives (FN) for a threshold = 0.2; and B) the number of False Negatives (FN) for a threshold = 0.6. Is it possible that:	☐	☐	☐

Task 4/4 - Questionnaire p.4/4

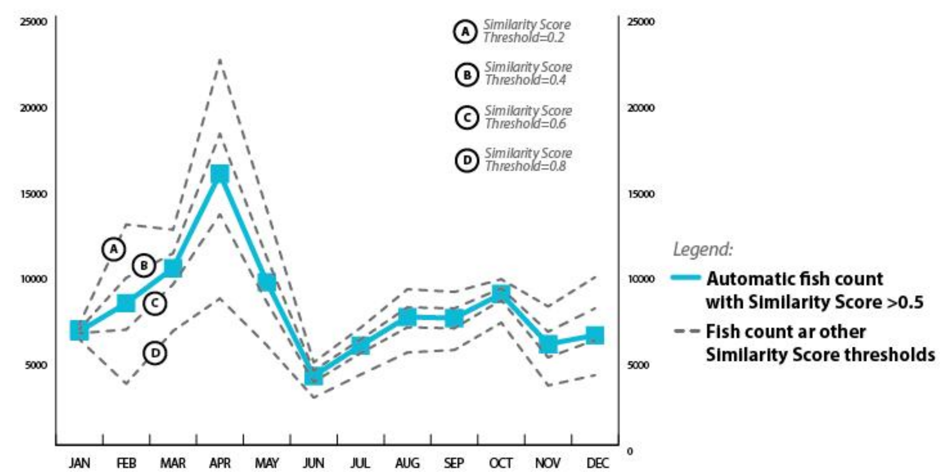
***31. The image below shows the counts of fish obtained by using different certainty score threshold. Which is the most appropriate thershold?**

- ○ Threshold A
- ○ Threshold B
- ○ Threshold C
- ○ Threshold D

Why did you choose this thershold?



The image below show the fish counts for the year 2011, as calculated using difference Similarity score thresholds.



***32. Which fish count would you choose to use for studying the variation of fish count over time?**

- I would choose the automatic count A (the upper dashed grey line).
- I would choose the automatic count B (the 2nd dashed grey line).
- I would choose the automatic count with a threshold of 0.5 (the blue line above).
- I would choose the automatic count C (the 3rd dashed grey line).
- I would choose the automatic count D (the lower dashed grey line).
- I don't know.

Why would you choose these fish counts?

＊33. Is there other source of errors?

The evaluation of the video analysis errors is made for a small set of Videos for Evaluation (as shown in the "Video" tab). But the video analysis software is applied to a much larger set of videos collected for the year 2011. When analyzing a larger set of videos, do you think that it is possible to encounter the following sources of errors. And would you like to measure and verify these possible errors.

	Is it possible these errors	Can it cause the automatic count to be different from the expert count?
The automatic fish counts is likely to contain non-fish objects (e.g., rocks) that are incorrectly considered as being a fish.		
When one single fish swims in and out of the camera's field of view, it is counted several times by the video analysis software. It is as well counted several times in the manual fish count.		
The automatic fish counts is likely to miss some fish that are not detected at all.		

Do you think of any other reason that could explain why the automatic counts are different from the experts' count?

Appendix B Goal-driven Task Model Used for Evaluation of Awareness

1. **GOAL:Decide whether there is a relation in species occurrence**
   (a) SUBGOAL: Find a pattern in species occurrence
      i. *DECISION:Is there a relation in occurrence of species X,Y and Z?*
         A. Is there a relation in occurrence of species X and Y? (*Projection*)
         B. Is the abundance of species X lower than species Y? (*Comprehension*)
         C. What is the number of fish of species X?(*Perceprion*)
   (b) SUBGOAL:Decide whether the observed pattern depends on time of year.
      i. *DECISION:Which period of year has the most fish?*
         A. What is the period of year we can observe the most abundant fish population? (*Projection*)
         B. What period of year has the most fish? (*Comprehension*)
         C. What is the number of fish occurring during the specific week of year? (*Perception*)
      ii. *DECISION:Is the relation in species occurrence depends on time of year?*
         A. Is a relation in occurrence of species X and Y depends on periods of year? (*Projection*)
         B. Is the abundance of species X lower than species Y at this time of year? (*Comprehension*)
         C. What is the fish counts of species X at this time of year? (*Perception*)
   (c) SUBGOAL: Decide whether the observed pattern depends on a location.
      i. *Decision: Whether an observed fish occurrence is specific to a particular habitat?*
         A. Is the pattern will occur in different locations with similar conditions(e.g. Camera 37, NPP-3)? (*Projection*)
         B. is the fish pattern for the specific period at location X is similar to location Y? (*Comprehension*)
         C. What is the number of fish occurring during the specific week of year at location X? (*Perception*)
         D. For which camera are we counting fish? (*Perception*)
      ii. *DECISION:Is the relation in species occurrence depends on a location?*
         A. Is a relation in occurrence of species X and Y depends on the habitat? (*Projection*)
         B. Is the abundance of species X lower than species Y at this particular location? (*Comprehension*)
         C. At which is the location we can observe the most abundant population of species X? (*Comprehension*)
         D. What is the fish counts of species X at this location? (*Perception*)
         E. What is the most abundant species at particular location? (*Perception*)
   (d) SUBGOAL: Validate the pattern
      i. *DECISION:Is the observed pattern introduced by the selection of video samples?*
         A. Whether the analysed number of videos suitable for scientific analysis?(*Projection*)
         B. Is the number of video samples constant over the analysed period?(*Comprehension*)

C. Is the number of video samples constant over the analyzed locations?(*Comprehension*)

D. How many video samples were analysed at the specific period?(*Perception*)

E. How many video samples were analysed at the specific location?(*Perception*)

ii. *DECISION:Is the observed pattern introduced by the performance of the software?*

A. Is pattern could be caused by the fact that the software does not detect fish accurately at this period of time?(*Projection*)

B. Is pattern could be caused by the fact that the software does not detect fish accurately at this location?(*Projection*)

C. Is the distribution of certainty scores different for another period of time?(*Comprehension*)

D. Is the distribution of certainty scores different for another Location?(*Comprehension*)

E. What is the distribution of certainty scores for the observed dataset?(*Comprehension*)

F. What is the number of fish with certainty score more than the selected threshold? *Perception*

Appendix C Tasks Used to Measure Awareness

***1. What is the number of fish occurring during Week 12?**

***2. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

***3. For which cameras are we counting the fish?**

***4. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

***5. Which week of the year has the most fish?**

***6. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

***7. At which period of the year can we observe the most important fish counts?**

***8. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

***9. How many videos were analyzed for the Week 12?**

***10. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

***11. What is the average number of fish per video for the Week 12?**

***12. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

***13. What is the fish abundance for the Week 45?**

***14. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

***15. Which week of the year has the most fish per video?**

***16. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

***17. At which period of the year can we observe the most abundant fish population?**

***18. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

***19. Is it the same for the fish population occurring at Camera 37?**

***20. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

NEXT SITUATION

SITUATION B

***21. Is the number of video samples constant over Hours of the Day?**

- Yes
- No

***22. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-					

If you have any comment, please let us know:

***23. Is the number of video samples constant over Weeks of the Year?**

- Yes
- No

***24. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-					

If you have any comment, please let us know:

***25. Is the amount of video samples suitable for scientific data analysis?**

- Yes
- No

***26. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-					

If you have any comment, please let us know:

***27. Which is the most abundant species in HoBiHu?**

- 1 Dascyllus Reticulatus
- 2 Chromis Margaritifer
- 3 Plectrogly-Phidodon dickii
- 4 Acanthurus nigrofuscus
- 5 Myripristis berndti
- 6 Chaetodon Trifascialis
- 7 Zebrasoma Scopas
- 8 Scolopsis Bilineate
- 9 Amphiprion Clarkii
- 10 Siganus Fuscescens
- 11 Pomacentrus amboinensis
- 13 Scaridea
- 26 Abudefduf vaigiensis
- 27 Lutjanus fulvus
- 36 Arothron hispidus

***28. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

***29. For which camera can we observe the most abundant population of fish from species Chromis Margaritifer?**

***30. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

***31. Do fish from Species Chromis Margaritifer generally have high certainty scores?**

- Yes
- No

***32. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

***33. Is the abundance of species Chromis Margaritifer lower than species Dascillus Reticulatus?**

- Yes
- No

***34. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

***35. Is this due to the fact that the video analysis does not accurately detect the species Chromis Margaritifer?**

- Yes
- No

***36. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

NEXT SITUATION

SITUATION C

Please obtain visualizations that supply the following information:
Considering all time periods and locations, what is the relative abundance of fish from each species (e.g., Species Composition) over the Weeks of the Year?

***37. Is there a relation in the occurrence of fish from Species 9, 26 and 27 over Weeks of the Year?**

***38. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

Please obtain visualizations that supply the following information:
Considering all time periods and locations, what is the relative abundance of fish from each species (e.g., Species Composition) over the Hours of the Day?

***39. Is there a relation in the occurrence of fish from Species 9, 26 and 27 over Hours of the Day?**

***40. What is your level of confidence in the above answer?**

	Very Low	Low	Moderate	High	Very High
-	○	○	○	○	○

If you have any comment, please let us know:

Appendix D Description of the Situations Used to Measure Awareness

Situation A Users mainly need to read the predefined visualization. They also have to manipulate the Y axis, to open a widget to fulfill, and to change its filtering parameters. 3 widgets are opened, all of them being irrelevant.
Complexity: 1/3
Objective: reading simple graph, understanding fish per video, manipulating filters, manipulating Y axis, recovering from irrelevant widget
Initial settings of the User Interface
y-axis: Fish Count
x-axis: Week of Year
Decomposition: none
Filters pre-selected: Year: 2011, Hour:All, Week: All, Camera: 38, Species:All, Certainty: [0,1], Comp.:D50-R52
Opened filtering widgets: Species, Component, Certainty

Situation B Users need to manipulate the X & Y axes of the graph, and to modify a few filters (only 1 in the optimal case). Some filters provide answers to the questions through their histograms. There is an alternative to reading the histograms: using several variations of the main visualization. This is longer, and is not an optimal choice. A few filters are opened, all being relevant. A lot of filters parameters are preselected (more specific data set).
Complexity: 2/3
Objective: manipulating X & Y axes, reading filters' histograms, reading predefined filters, open relevant filter, choose between using filters or visualization **Initial settings of the user interface**
y-axis: Video count
x-axis: Hour of Day
Decomposition: none
Filters pre-selected: Year: 2011, Hour:All, Week: All, Camera: 40, 41, 43, Species:Chromis Margaritifer, Certainty: [0,1], Comp.:D50-R52
Opened Filtering widgets: Camera, Species

Situation C Users need to manipulate the X & Z axes of the graph to fulfill the task. Some filters parameters are pre-selected, most of them being irrelevant. Only 1 filter is opened, and is irrelevant.
Complexity: 3/3
Objective: manipulating X & Z axes, recovering from irrelevant filtering especially when widgets are hidden.
Initial settings of the user interface
y-axis: Fish Count
x-axis: Week of the Year
Decomposition: none
Filters pre-selected: Year: All, Hour:11am, Week: All, Camera: 38, Species:2, 4, Certainty: [0,1], Comp.:D50-R52
Opened filtering widgets: Certainty

Appendix E Description of the User Interface Addressing Data Provenance Needs

The Video tab

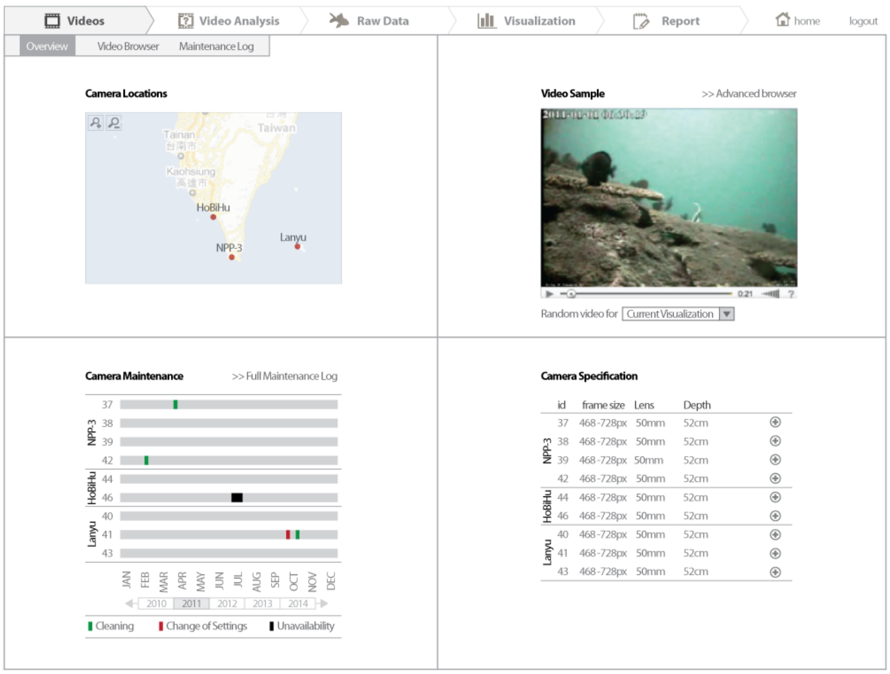


Figure 11: Video tab. Overview of technical characteristics of the video collection

The Video tab provides an advanced video browser and access to maintenance logs and camera specifications. The content is organized in 3 sub-tabs. The 1st sub-tab (Fig. 12) provides a overview of the entire characteristics of the video collection: the location of camera, the video images, the important maintenance operations, the camera specification (e.g., resolution, lens, depth in the water). The 2nd sub-tab (Video Browser in Fig.13) provides an advanced video browser that can retrieve videos containing specific species or behaviours. The last sub-tab (Maintenance Log in Fig.14) provides both a tool to browse and retrieve maintenance operations, but also to edit the logs (e.g., add, delete, modify logs). We assume that biologists are likely to be aware of the maintenance operations, or other events happening to the cameras. We assume that they currently have no mean to record and centralize these events. This is why the User Interface supports the functionalities to add, delete and modify the maintenance operation logs.

The Video Analysis tab

The Video tab provides advanced functionalities to explore the ROC evaluation of the Fish4Knowledge video analysis components. The content is organized in 4 sub-tabs. The sub-tab "Overview" (Fig. ??) describes the video analysis process step by step. The sub-tab "Fish Detection" (Fig. 15) provides ROC explanations for each version of the Fish Detection component. It also reports ROC evaluation of the algorithm over different water conditions. The sub-tab "Species Recognition" (Fig. 16) provides the ROC explanations for each version of the Species Recognition component.

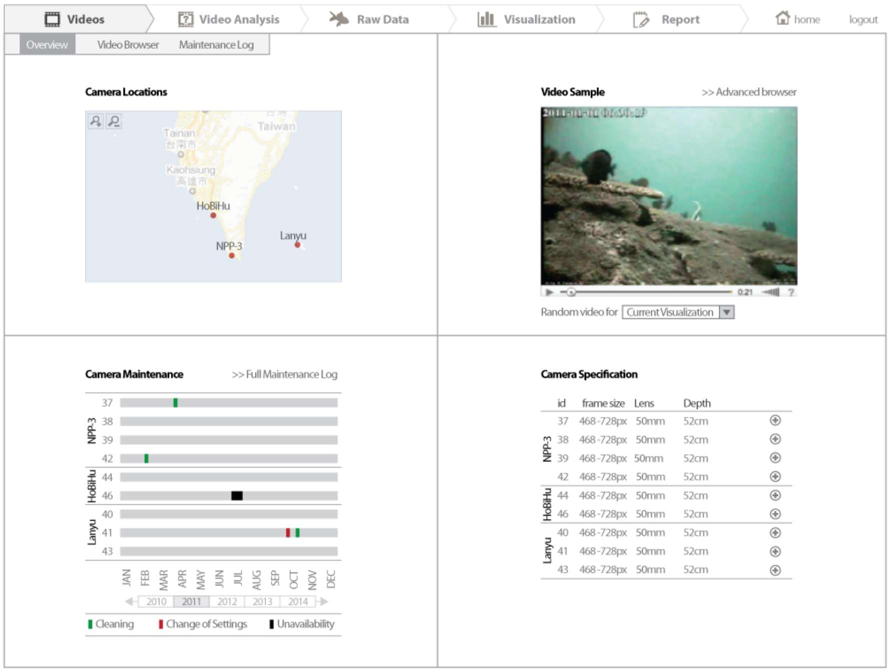


Figure 12: Video tab. Overview sub-tab

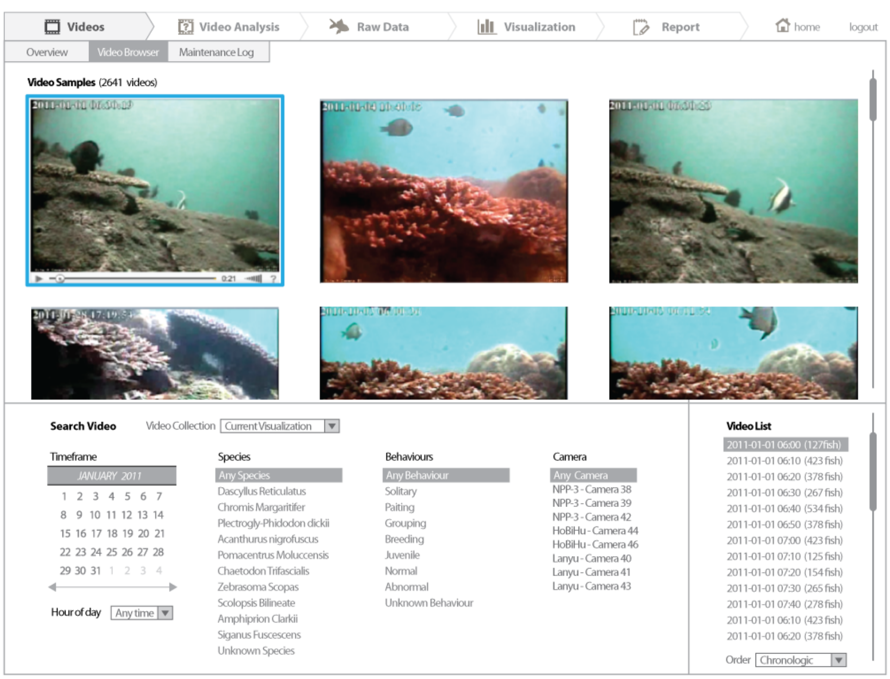


Figure 13: Video browser

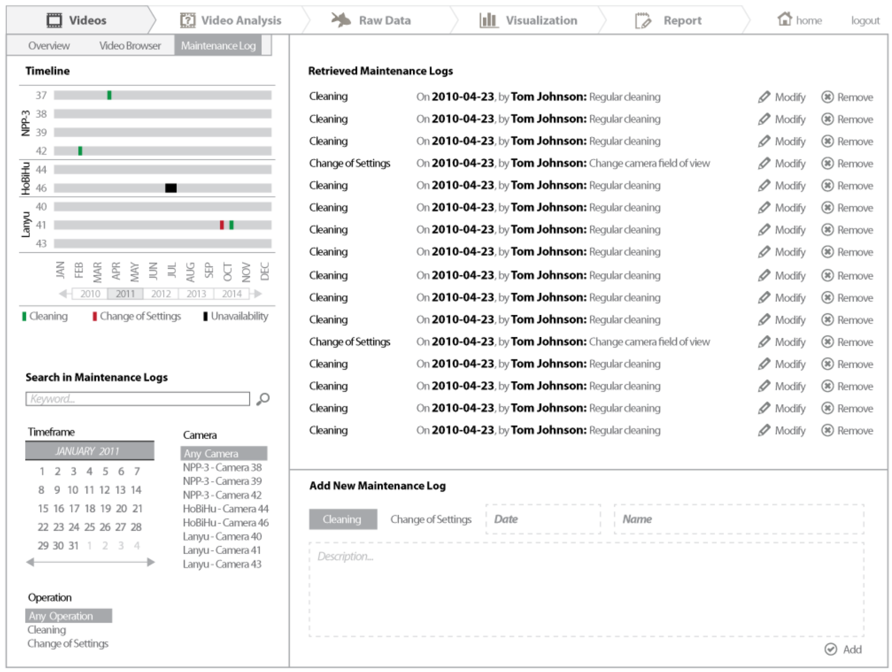


Figure 14: Video Maintenance

It also shows the evaluation of algorithms for each species in particular and risk of confusing the species with each other. The sub-tab ”Workflow” (Fig. 17) provides a tool to request a video analysis task of a specific dataset.

The Raw Data tab

The Raw Data tab provides an overview of the available data, and the characteristics that can be used for filtering. It includes video data that can be used to evaluate the potential errors and uncertainties, the processed and erroneous videos, the versions of the video analysis components, the similarity score thresholds, and the number of items in the ground-truth. The content is organized in 3 panels as shown in Fig.18. The top panel provides a graph that represents the structure of the Fish4Knowledge data, from a user point of view. The boxes represent the entities that can be counted, or used to decompose other counts of entities (e.g., counts of fish over species). It also represents the characteristics of the data that can be filtered. The contours of boxes indicate the type of parameters used for filtering: custom filters with manually-defined parameter values, or default filters with automatically-defined parameter values. The default parameters concern the similarity thresholds, the versions of the software components (as decided by the Workflow), and the species and behaviour (all of them). The middle and bottom panels are the same as for the Visualization tab. They contain the functionalities for filtering the data (bottom panel), and for specifying the visualization of interest (middle panel).

The Visualization tab

The content is organized in 3 panels as shown in Fig.19. The top panel consists of a graph that can be used to visualize counts (of fish, species, videos, ground-truth items, or versions of components). The counts are represented on the y-axis. The x-axis can represent various dimensions of the data,

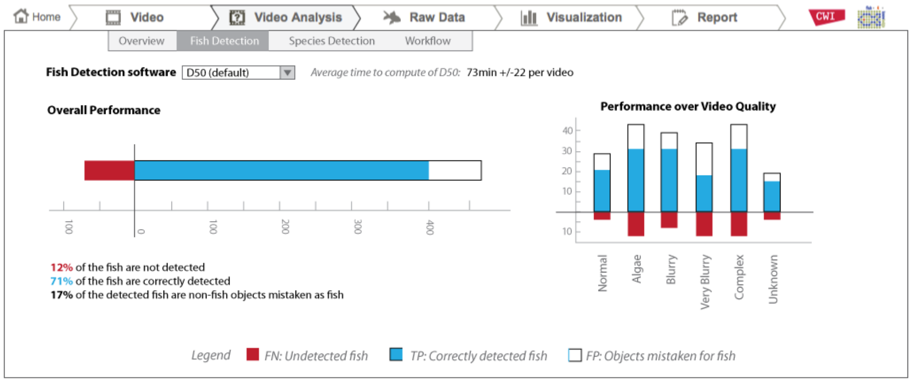


Figure 15: Video Analysis tab. Fish Detection sub-tab

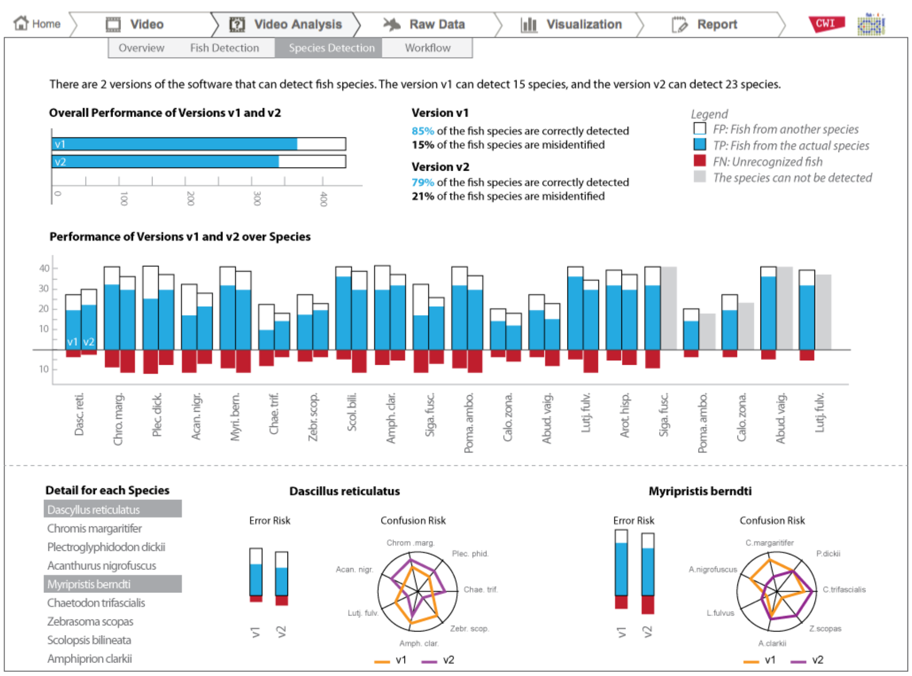


Figure 16: Video Analysis tab. Species Recognition sub-tab

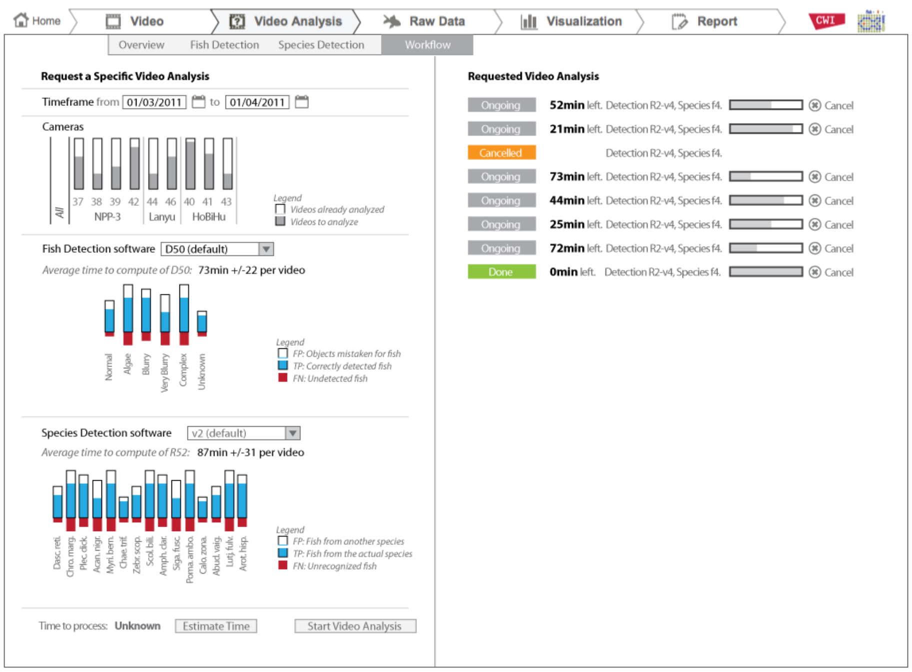


Figure 17: Video Analysis tab. Workflow sub-tab

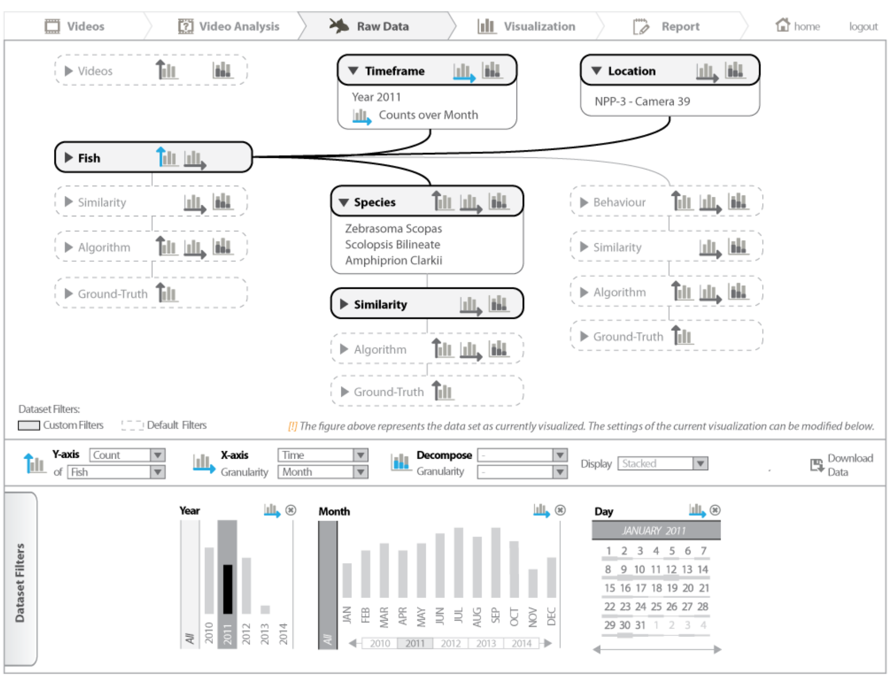


Figure 18: Raw Data tab

Figure 19: Visualization tab. Visualization of fish counts over weeks of year

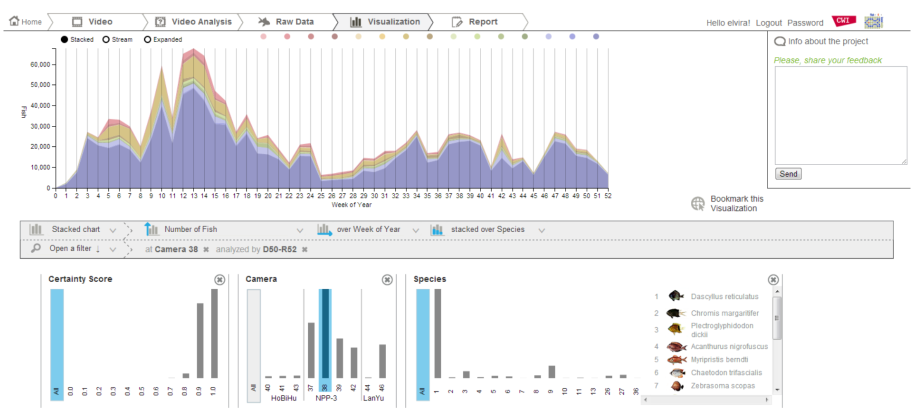


Figure 20: Visualization tab. Fish counts distributed over weeks of year and decomposed by species

depending on what users selected for the y- axis. For instance users can specify the following visualizations: fish counts on the y-axis and list of species on the x-axis; counts of species on the y-axis and week of the year on the x-axis; counts of processed videos on the y-axis and cameras on the x-axis. Additionally, the graphic supports advanced functionalities to visualize data, and the decomposition of counts. The user chooses to decompose the fish counts per species, and display them in a stacked graph as shown in Fig.20. This allows biologists to study species composition.
In the middle panel, a user can interact with the drop-down menus to specify the visualization of interest. The first drop-down menu contains the types of visualizations that are available in the Fish4Knowledge interface (e.g. stacked chart). In total, there are 3 types of graphical representation available: *error bars* or a *box plot*, *single data points*, *stacked counts*. A *box plot* is a standard visualization used by a biology community for studying the statistical variability. *Stacked chart* allows biologists to study the composition of objects over one extra dimension. Each metric calculated for the y-axis (count or growth rate) over each bin of the x-axis is decomposed into a set of metrics calculated for smaller samples (e.g. decomposition of fish counts over weeks of year, decomposition of fish counts over cameras). The second and third drop down menus allows users to select x- and y-axis respectively. The options in these drop-down menus are determined by the choice of the graph made in the first drop-down menu. The *y-axis* can be counts of either: fish, species,behaviours, videos, ground-truth items, or versions of components. The *x-axis* can be time periods, cameras, species, behaviours, similarity scores, versions of components.
In case if a biologist selects a box plot or a stacked chart as a representation type, the forth drop-down menu appears and allows to specify grouping parameter (e.g. week of year, species). The grouping parameter for a box plot could be time periods. The grouping parameter for a *Stacked chart*(as a form of z-axis) could be species, cameras.
The second row in the middle panel allows users to work with filtering widgets. It also provides an overview of the filters applied to the currents dataset. The first drop-down menu shows the list of available filtering widgets . The data can be filtered on the following dimensions: time periods(year, week of year, hour of day), cameras, certainty scores, species and software components. For instance, the users could visualize the fish counts for one particular species only for the winter months by specifying the particular weeks in the *Week of Year* widget. All the filtering widgets contain a small data visualization in the form of histograms. It represents the visualization that could be obtained in the data used for filtering was also used as the x-axis of the main graph of the top panel.

The Report tab

The Report tab provides a simple mean to group and comment visualizations of interest (Fig. 21). Users can add, delete, modify the visualizations, the titles and the comments. They can also download the reports, with the possibility to attach the raw data of the visualizations, and the evaluation of the underlying video analysis processes (e.g., ROC evaluation of the components involved).
Users can add visualizations in the report from the Visualization tab, by using a button on the right side of the middle panel. Reports can be saved and continued later by using the Download Report and Upload Report functionalities. Users can locally store a report, or share it with their colleague as a regular file (e.g., to be sent by mail). Reports can be uploaded and users can visualize or modify them. This offers basic but limited support to collaborative and long-term studies of the Fish4Knowledge data. It avoids the engineering problem of storing and sharing users data analyses online. For instance it would involve the implementation of a user right management system, for which we can not support the high implementation cost.

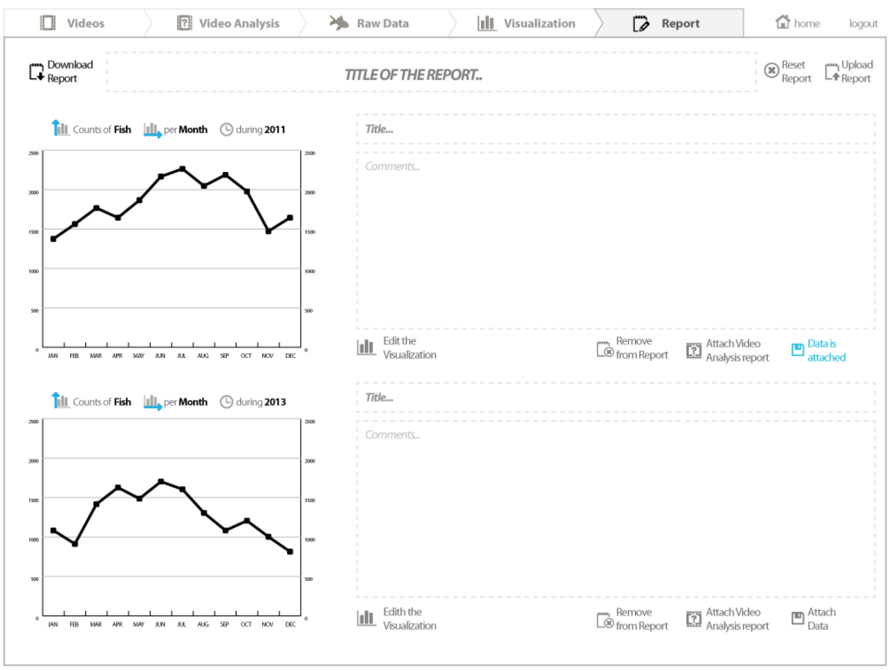


Figure 21: Report tab