Fish4Knowledge Deliverable D5.3

Scientific Question and Experiment Plan

Principal Author: J. van Ossenbruggen, L. Hardman, E. Beauxis-Aussalet, J. He, B. Boom, C. Spampinato

Contributors: CWI, UCAT, UEDIN

Dissemination: PU

Abstract: This document specifies scientific questions relevant to evaluating components of the F4K system, in addition to evaluating the system as a whole.

Deliverable due: Month 3

Contents

1 Introduction

The F4K system consists of a number of components that work together by storing information into a shared data repository and/or retrieving data from this repository. Tools are developed within the project to:

- Analyse video data, and store the results into a dynamically growing database,
- Perform data enrichment strategies to further enrich the database
- Construct user interfaces to provide access to this database for marine biologists.

The purpose of these tools is to allow marine biologists to find answers to questions such as those described in D2.1 User Information Needs.

This document describes scientific questions applicable to each of the components as well as an evaluation strategy for these and the system as a whole. For the overall project, we identify the following general research questions:

- Can we develop technology that is able to detect and track individual fish from real life video footage?
- Can we develop technology that is able to recognize the species, genus or family of the fish detected?
- Can we develop technology to make the results produced by detection, tracking and recognition accessible to marine biologists who are not experts in computer vision?
- Can we do all of the above in a way that allows biologists to judge to what extent the presented results are sufficiently reliable to be used in scientific publications?

These research questions reflect some of the key areas of innovation of the F4K project.

First, the detection, tracking and recognition software will be deployed on data that has not been recorded in a controlled laboratory environment, but by cameras located on real coral reefs in open oceanic environments. This implies a) that the software needs to be robust, and b) needs to scale to the large amounts of video footage that are generated by the project's cameras on a daily basis.

Second, if we would only allow direct access to the raw data produced by the detection, tracking and recognition software, we run the risk that it will need to be interpreted by users that are both experts in computer vision and in marine biology. To make the data accessible to marine biologists, who are not experts in computer vision, requires the development of a user interface that *can hide technical details that are irrelevant to the current task* but at the same time can *guide non-expert users in how to access and interpret these technical details and how they influence the conclusions being drawn.*

Finally, to answer the final research question we need to identify the different roles of uncertainty in each component of the system, how these influence one another and what their effect is on the final outcome of the results presented in the interface. This will be a non-trivial process. In general, we assume that scientific users will only trust the system if every single result presented can be explained and its value verified by the user, e.g. for every result it should be clear from what data it was derived and how. Biologists should also be sure that the F4K system does not introduce barriers with respect to the replicability of the biologist's conclusions. On the one hand, it should be easy to rerun a previous analysis on new data, or by using improved versions of the detection software. On the other hand, it should also be possible to replicate a previous analysis under exactly the same conditions using exactly the same data and software, even if both data and software have evolved since the previous analysis. Finally, depending on the research question, the marine biologist may choose to assume that data presented by the F4K system are reliable and apply the usual statistical analysis to investigate these data support or reject a specific hypothesis. More likely however, the inherent uncertainty of the data presented by F4K will play a role in this analysis and will need to be taken into account.

Reference datasets are datasets that contain similar data to that produced by our software, but consist of data that is manually generated or has been manually checked for correctness by human experts. Such datasets play different roles in this project. First, they can be used to evaluate the quality of software components by comparing the results produced by the software to the reference data. Second, they can be used as training data in machine learning contexts. Note that once a system has used a dataset as training data, that same set can no longer be used for evaluation purposes. Finally, reference data can be used in the user interface as part of the explanation to end users of the internal workings of the system.

An initial ground truth dataset will be prepared, and published on the web, for at least 1 hour of video (1 minute at 3 times of day for 2 days in different seasons at each of the 10 cameras). The ground truth will contain the locations of fish, identification of interactions and classification of species. Since creating larger reference datasets is a very laborious and expensive task, the creation of such sets has become a specific research goal in itself. We first describe the evaluations that can be carried out on separate components of the system, after which we discuss end-to-end evaluations.

2 Fish detection and tracking

The first step of making processed data available and understandable for marine biologists is the detection and tracking of fish in real-life underwater unconstrained environment. Detection and tracking information together with colour and texture information will be then used to identify events of possible interest for marine biologists.

The first research question is:

- **Can we develop fish detection algorithms that are adequate in terms of accuracy and efficiency for a real oceanic environment in contrast to high quality images from a controlled laboratory environments?**

This research question will involve the investigation of fish detection approaches to handle effects that usually occur in the observed underwater scenes and that might affect image quality, such as the periodical gleaming on underwater scenes, the weather conditions (sudden cloudiness, storms and typhoons) that make hard to detect clearly any targets due to a worsening of image contrast, the murkiness of water and the algae and filth growing on camera lens due to the direct contact with seawater.

One important aspect of detection algorithms that will be taken into consideration is the periodic and multimodal nature of background of underwater scenarios; in fact, handling background movements and variations is one of the most difficult tasks and detection algorithms must be robust enough to cope with any arbitrary changes in the scene. Also periodic movements (e.g. plants affected by flood-tide and drift) have to be taken into account to avoid the detection of moving non-fish objects.

A key challenge of our low level video analysis will be to design object segmentation approaches, able to perform well with low resolution images, in order to improve the quality of the masks extracted by the detection algorithms. This is heavily demanded since most of the recognition algorithms will rely on the extracted fish contours/masks.

On the other hand, the output of the recognition step (see next section) will be integrated with the detection algorithms (feedback) to improve their performance, more specifically to reduce the number of false positives that may occur during the detection phase.

However, one of the most complex issues to deal with, when processing underwater videos, concerns the fish tracking step. This leads to our second research question:

- **Can we develop tracking algorithms that can track reliably the same fish over multiple frames?**

Differently from humans, fish perform erratic and fast movements (in three dimensions) resulting in frequent changes of size and appearance. The tracking modules have to adapt the model effectively both to scene and fish appearance variations.

The tracking step is necessary to: 1) provide precise measurements of fish counting that, instead, cannot be derived by using only detection approaches, 2) handle and correct errors occurred during the detection, as for instance occlusions among fish, and 3) to extract reliably fish trajectories for the subsequent trajectory-based behaviour understanding.

Both detection and tracking tasks are, however, influenced by the frame rate of the recorded videos since most of them rely on modelling a state representing the objects position and appearance in consecutive frames. An analysis on how frame rate influences detection and tracking results will be carried out.

The last research question is:

- **Can we develop robust approaches to identify events of potential interest to marine biologists?**

The event detection part relies on the previous detection and tracking tasks. Assuming a correct working of these modules, we will investigate how to detect fish activities and fish-fish and fish-background interactions, by exploiting different cues like trajectory analysis, changes of appearance and changes of colour and texture patterns. Time variations of these features will be combined using inference to detect events of interest. Contextual information such as background objects' position and event hierarchical definition will be also included in the analysis.

Since event detection involves other information (i.e. colour and texture patterns) than trajectories, object description techniques invariant to affine image transformation will be also explored.

Answers to these three questions are expected to be publishable in vision conferences such as ICIP, ICPR, ICCV, CVPR and related journals IEEE TIP, IEEE PAMI, IJCV and in multimedia conferences ICMR, CBMI, ACMMM and related journals IEEE Multimedia, Springer MTA, ACM TOMCCAP.

2.1 Evaluation Measures

2.1.1 Detection and Tracking Algorithms

As highlighted above, these tasks aim, respectively, at detecting and tracking correctly fish. To evaluate the effectiveness of these approaches reference data is heavily demanded. There exist a lot of reference data for testing detection and tracking approaches, especially the ones of the earliest versions of the PETS workshops that included datasets for surveillance of public spaces, detection of left luggage, outdoor and indoor people tracking. Other reference data are the CAVIAR, i-LIDS and ETISEO data sets. Therefore, we will test our approaches on the above reference data, but since we deal with underwater environment and there no exists reference data in such scenario (or in scenarios with similar peculiarities) we will create a reference dataset for low quality unconstrained environments.

For testing the detection algorithms it is necessary to evaluate 1) the ability in detecting objects through object-based metrics and 2) the quality of the extracted masks through pixel-based metrics using a reference data where objects of interest are identified and segmented. In both cases, typical metrics, computed per-frame or per-sequence, are: the true positive rate (or detection rate), false positive rate, false alarm rate and specificity. For object-based evaluation it is not possible to estimate the false positive rate since it is not possible to identify true negatives.

Since these metrics assess overall segmentation quality on a frame-by-frame analysis, but fail to provide an evaluation of individual object segmentation, we will normalise these measures by image size, amount of detected change in the mask and object relevance. The common Receiver Operating Curve (ROC) will be also used.

For testing the tracking algorithms we will use, instead, tracker detection rate, false alarm rate and object tracking error. Furthermore, since in the case of tracking it is not straightforward to say whether an identified trajectory is right or wrong when compared to a reference trajectory (the two might overlap only partially), we will test our approaches also using self-evaluation based approaches, which assign a score to trajectories according to criteria of regularity in direction, speed, appearance of the extracted tracks.

2.1.2 Event Detection

The event detection part aims at identifying interesting events for marine biologists; also in this case there no exists reference data for testing event detection approaches in underwater footage and we will build our own reference data containing fish events.

However, one reference point for event detection research community is the TRECVID workshop series, more specifically, the Multimedia Event Detection (MED) tracks where, in the last few years, reference data for detection of events involving people has been provided.

We plan to test our approaches on our fish reference data; and the applicability of such approaches to other domains (e.g. people, sport, etc.) will be investigated through the participation to one of the forthcoming TRECVID workshops (likely the one in 2013).

The primary evaluation metric of event detection approaches is the actual normalized detection cost (NDC) that has to be computed for each event independently. The evaluation results for the detection of the target events will be also performed using the detection error trade-off (DET) curve and the ROC curve for each event.

2.1.3 Performance measures

The performance of the detection and tracking algorithm will be measured in terms of processed frames per second and memory occupied; whereas the performance of the event detection module will be given by the time to detect a specific event in a standard video sequence. Scalability is also an important issue of event detection approaches.

3 Fish species detection

Based on the detection and tracking modules described above, raw numbers of fish per period and location can already be calculated. To make the data even more useful for marine biologists, we will run automatic species recognition on the data produced by the detection and tracking.

3.1 Recognition

Our key research question is:

- Can we develop recognition algorithms that can identify the species of an individual or cluster of detected fish?

Since the recognition runs on the output of the detection and tracking, an important derived question is how the quality of that data influences the recognition quality. Special attention will be given to false positives in the detection, segmentation errors and occlusion. Additional challenges arise from the skewed distribution of the various species in the observed data, e.g. few species occur very frequently, while most others are quite rare, sometimes so rare that it will be impossible to obtain sufficient training and evaluation data for these species.

As discussed above, reference data will be needed both to train and to evaluate recognition techniques. As reference data for this task is rare, we are considering testing our approach on alternative, but closely related tasks. For example flower and bird recognition standard data sets such as the Oxford flower database and Caltech-UCSD Birds 200 are available. For fish recognition the key reference dataset is fishbase.org, which typically consists of still images taken in a controlled environment or even drawing. Creating a reference dataset that is representative for the real life footage produced in this project will thus be a research objective in itself, as discussed in the next section. Here it suffices to state that, due to the large numbers of detection in the database it will be a necessary preprocessing step to cluster similar fishes prior to the species recognition, that is, to cluster individual fishes that are likely to belong to similar species, purely on visual characteristics. Our second research question is thus:

- Can we develop clustering algorithms that can group detected fish of the same or similar species together?

Given the fact that the clustering is used to create the evaluation data for the recognition, special care should be taken to avoid or minimize the chances that limitations of the clustering bias or otherwise negatively affect the quality of the final reference data.

3.2 Evaluation measures

Recognition quality The main task of the fish recognition is to recognise fish correctly. The commonly used evaluation for recognition methods is the receiver operating characteristic (ROC), which plots the false positive rate against the true positive rate by shifting the discrimination threshold. A simpler measure is the recognition rate, which tells how much fish in the database are recognised correctly. Because we have classes with different distribution, a mean class recognition rate will be used in these cases. Finally, we can also consider the precision versus recall (PR) curves, because of the large amount of data, we discussed this further in the next chapter.

Clustering quality The main task of the fish clustering is given a fish image to retrieve images of similar fish species. In literature, this is also known as (content-based) image retrieval, where we use an image as query. The common measure to evaluate the retrieval results in precision versus recall (PR) curves. Related measures used in this case are the average precision, which approximates the area under the precision-recall curves. For clustering, we can also use the same measures as the recognition given we have to match an example image against a larger set.

Performance measures One of the important other aspects of the fish clustering method should be the speed with which it is able to retrieve the fish images. Scalability is an important property of the clustering method since it should be able to deal with large amount of data. The most common measurement is in these cases the time it takes to cluster a given amount of images.

4 Obtaining reference data by crowdsourcing

High quality reference data are necessary for both evaluation and development purposes, but obtaining this on a large scale is a tedious and expensive task. We intend to achieve this through crowdsourcing, where the data annotation task is distributed to large numbers of anonymous annotators (the workers) on the Web. Using crowdsourcing requires knowledge on ways to encourage workers to want to and to be able to carry out tasks, while at the same time ensuring high quality output of the task. For a baseline crowd sourcing system, we will employ a standard crowd sourcing platform, namely CrowdFlower, which allows European users to create jobs and access the Mechanical Turk platform, which is the largest crowdsourcing platform on the Internet.

Two basic types of ground truth are required: firstly, whether the shape detected is a fish and if so which images belong to the same fish, and secondly, to which species, family or genus the detected fish belongs. Our research questions are thus:

- Can we use crowdsourcing techniques to reliably produce large sets of reference data for fish detection, tracking and recognition?
- How does the expertise of the users influence the quality of the results?
- How can we design the crowdsourcing tasks in a way that a minimal level of expertise can still yield results of sufficient quality?

- How can automatic methods improve the crowdsourcing process?

Incentives Workers should be motivated to work on the task. A widely used incentive is money, which is: 1) motivating for both tedious and enjoyable tasks; 2) easy to measure and can be flexibly adjusted for balancing cost-profit. An alternative incentive is entertainment, where workers will play for their own amusement. For example, the classification of fish images could be implemented as a Tetris style game.

Expertise required Non-expert workers can be recruited for identifying occurrences of fish and their continuous movement. The fish recognition task, however, requires specialist knowledge to identify the most likely species, genus or family to which the identified fish belongs. To allow crowdsourcing to help with these expert annotations, we can simplify the task to that of comparison with a pre-selected set of fish of which the species are known. For example, a worker would be presented with a test image and asked to select the most similar fish from a set of known species. The classification is based solely on visual similarities and no expert knowledge is required from the worker. To increase the number of identifications per user, some form of preprocessing the data in the form of clustering will be needed. We will research what approaches make the identification task as effective as possible, without biasing the resulting data set.

Quality control In order to obtain reliable annotations, defensive strategies against cheats and sloppy workers are necessary, particularly where money is used as an incentive. We consider the following basic defensive strategies:

1. Groundtruth seeding. Mixing ground truth with test data can identify low quality workers that frequently make mistakes on the ground truth data.
2. Redundancy. Present each image to multiple workers to allow measures of agreement to be formulated, such as the most often chosen species, or the most consistent workers.
3. Reputation. Crowd sourcing platforms, such as Mechanical Turk, provide a reputation system to motivate workers to take the task more seriously, e.g. to build up their reputation for easier access to future jobs.
4. Reviews. i) Multilevel review: use a set of workers to do the task and an independent second set of workers to evaluate the quality of the task results; ii) Expert review: Use a trusted expert to assess the accumulated for consistency.

Publications on these topics will appear in venues such as: SIGIR, CIKM, ECIR, WWW, ACM MM; and related journals: JASIST, JIR, IPM, TOIS.

5 Information access and interactive analysis

5.1 Understanding user needs

- What are marine biologists' requirements and information needs for understanding and trusting video analysis data, and for using these data in the context of their scientific research?

Supplying domain-specific information Video analysis components populate a repository of data (computer vision data) that describes the objects and events identified in the videos, as well as describing the video analysis processes and results (e.g., version of the components, certainty of object detection). Biologists seek to analyse biological concepts (e.g., fish abundance) that can be measured using computer vision data. Our first research question is thus:

- What are the biological concepts that are of interest to biologists and that can be measured using the derived computer vision data? We will interview marine biologists about the most important queries they would want the system to answer. Throughout this process, we must be aware that:
  - Some of the user queries might not be feasible, and will not be supported.
  - Other feasible queries of interest might not be mentioned during the interviews and might arise later, or when using prototypes of the F4K tool.
  - Some of the user queries might require additional computation to measure biological concepts that are not directly expressed in the repository of computer vision data (e.g., rate of increase of a specific fish population).
  - Some of the user queries might require enrichment of our computer vision data with external data sources or prior domain knowledge.

Given these user queries, we need to identify in what form users expect data in response to their queries. For instance, users might request a list of counts of fish for each month of the year, or a list of the increased rate of these counts, whereas the computer vision data repository contains only a large number of fish detections. Thus we must also identify the data transformations needed to derive these measures. To summarize, we have to identify:

- The domain-oriented data that are requested through user queries.
- The transformations of computer vision data that must be performed to supply the derived domain-oriented measures.
- The terms to use to allow biologists to understand what domain-specific information is carried by the data they are supplied with.

Trust and scientific requirements Several computer visions algorithms, and their specific parameters, are involved in the production of measures supplied to biologists. As biologists need to assess scientific research, they need to access, understand and trust the layers of computation that are used to produce the video data. Furthermore, biologists need to understand how limitations and parameters of algorithms can impact their data analysis results. Thus we must investigate what verifications and controls on video analysis algorithms are of interests for biologists because it can impact their scientific research results; in other words:

- What underlying computer vision processes need to be understood, trusted and controlled by users?
- What are the user requirements for Fish4Knowledge data to be suitable for scientific data analysis?

Particular attention is drawn to specific uncertainties contained in computer vision data, compared to uncertainties that are already encountered and taken into account by the biologist community when using their own sampling and data analysis methods.

These questions are investigated by interviewing marine biologists, and by reviewing the literature reporting biologists' methods, and data analysis methods, especially regarding uncertainty and trust issues. The specification of user requirements and information needs is to be reported in the deliverables 2.1. and 2.2., and will be reviewed by domain experts.

5.2 Supporting user needs

- What computing and user interaction means can be designed and implemented to support information needs and requirements for biologists to utilize computer vision data?

The previous research question leads to the identification of information needs and requirements for biologists to utilize computer vision data, as well as a high-level definition of the user tasks. Given these requirements, we must investigate means to support biologists with usable and controllable data, understandable information and human-computer interface that support their data analysis tasks.

To do so, we must investigate the ability of different user interface paradigms to support user needs and, particularly, their ability to support the exploration and the analysis of large amounts of data, and to convey domain-specific information. We also have to consider the interactions supplied by each user interface paradigm (e.g., type command, zoom in and out, use of map) and the implications in detailed user task flows that can be designed. We also consider means to simplify and optimize their tasks, for instance by introducing shortcuts or predefined data analysis templates.

Valid user interface paradigms must then be compared for us to select the paradigm, or combination of paradigms, that are the most appropriate w.r.t. the fulfilment of user needs, the usability of task flows, the HCI research progress, and the implementation feasibility. In other terms, we aim at answering the following questions:

- What user interface paradigms, and detailed user interactions, can support user requirements, informations needs and data analysis tasks?

- How appropriate are these paradigms for the novel data analysis facilities they offer to biologists, and for the human-computer interaction research they support?

Relevant user interface techniques will include those for handling multi-faceted data analysis, large amounts of data and exploratory data visualization. Where appropriate, we will consider means to integrate prior domain knowledge, as well as the personalization of data analyses and rendering of results.

To decide on which user interface functionalities to implement, a trade-off will be found between implementation costs, and contributions to biology and human-computer interaction research.

The answer to this question is obtained by:

- Designing means to address requirements and information needs, and selecting solutions to implement by evaluating their relevance and feasibility,

- Analyzing how domain experts work to solve tasks using prototypes of the Fish4Knowledge system, or user interfaces dedicated to a specific facet to study,
- Reviewing selected literature in the fields of human computer interaction, visual analytics, knowledge discovery, information retrieval and design science.

Publications on these topics will appear in venues such as: IUI, SIGCHI, UIST, HCII, VAST, VDA, VL/HCC; and related journals such as: TOCHI, TOG, IJVAAIM, IJDMMM, IEEE CG&A.

6 End-to-end scalability evaluation

The general research questions mentioned in the beginning each have their counterpart in terms of end-to-end scalability evaluation.

6.1 Speed of detection, tracking and recognition at production site

To be able to process all newly incoming videos and to process the backlog of archived videos, the detection, tracking and recognition modules need to be able to process videos and store the results in the common data store at sufficient speed. This speed should significantly exceed the speed of new incoming videos. Based on 10 live cameras, the overall throughput should preferably be above 15 hours of video processed per hour, on the cluster facilities provided by NCHC. We will monitor the speed of processing once the software has been installed at NCHC. Note that sufficient speed is necessary to be able to fill the database with sufficient data. We do not plan a separate scalability test in a artificial setting, but opt to provide detailed statistics on the speed of the production environment.

6.2 User interface responsiveness on large datasets

The user interface needs to be able to access the produced data at a speed that is sufficient for interactive response. We will test the retrieval speed of all the queries related to the questions and tasks identified in deliverable D2.2, using at least an amount of data that is equivalent to 15 years of video being analyzed, or more if available. Interactive response times should be in the subsecond range for frequent queries and routine data analysis, and should gracefully degrade for more complex and infrequent queries.

6.3 End-to-end functionality & usability evaluation

Finally, the overall system needs to be evaluated by the primary target users: marine biologists. Since marine biology is in many aspects a location-specific research field, our main group of test users will be recruited from the marine biology research groups in the Taiwan region. Additional user testing will be done with participation of biologists from other areas, focusing on groups that have a direct link with the project, for example via the project's Advisory Board. User testing will be done using the data in the production database hosted at NCHC using the scenarios and test tasks identified in Deliverable 2.2.

Tasks will include the identification of the subset of the data that is relevant to the topic under study, analysis of this subset, visualization of the analysis results, and an assessment of the reliability of the results that take into account the uncertainties introduced by the F4K system.