Automatic Indexing of Underwater Survey Video: Algorithm and Benchmarking Method

Katia Lebart, Chris Smith, Emanuele Trucco, and David M. Lane

***Abstract*—It is often the case that only a few sparse sequences of long videos from scientific underwater surveys actually contain important information for the expert. Locating such sequences is time consuming and tedious. A system that automatically detects those critical parts, online or during post-mission tape analysis, would alleviate the expert workload and improve data exploitation. In this paper, a methodology for evaluating the performance of such a system on real data is presented. Interesting sequences are started by changes of visual context. An algorithm to detect significant context changes in benthic videos in real time has been presented by Lebart *et al.* in 2000. It is used as an illustration for this methodology—its performance is studied and benchmarked on real underwater data, ground truthed by an expert biologist. Various issues relating to the complexity of the problems of automatically analyzing underwater video are also discussed.**

***Index Terms*—Change detection, image analysis, indexing, light compensation, online image processing, underwater, unsupervised clustering, video.**

I. Introduction

VIDEO images are of paramount importance in underwater scientific missions for applications such as monitoring sea life, taking census of populations, and assessing geological or biological environments. Scientific underwater surveys produce a wealth of video data, but its exploitation is far from optimal. Retrieving the sparse relevant information from many hours of video produced by a mission is time consuming and tedious. Currently, experts have to either monitor the camera output during the mission or browse through the videos after the mission and log the interesting occurrences manually. Software packages are available for the logging of multidata and video online during remotely operated vehicles (ROV) operations [2]. For video, this takes the form of screen grabs with manual annotations. It is neither automated nor fully comprehensive, relying on the slow speed of operation and availability of the operator to make entries. Automated indexing of the video data would dramatically reduce the time required to extract the key frames of the video, leaving the expert to concentrate on the analysis of the data rather than on the data mining of the information. Along with the development of new algorithms, a suitable benchmarking framework must be established.

Progress in video-digitization hardware and digital cameras has recently enabled much research toward the underwater applications of computer vision. Most applications, however, address control tasks such as vehicle stabilization ([3]–[6]), cable/pipeline following ([7]–[11]), image mosaicing ([12]–[18]), navigation ([19]–[21]), and concurrent mapping and localization ([22]–[24]). Many of these techniques rely on some estimation of the apparent motion between successive images, either through gradient or optical flow estimation ([25]–[28]) or through characteristic point tracking ([29]–[31]) and [32] in the case of pipelines.

Very little work exists in statistical image analysis and pattern recognition. Some recent studies address constrained problems: the detection of manmade objects such as pipelines ([33]–[40]) or of military targets ([41]). Developments have been made for the classification of plankton ([42], [43]), the exploration of manganese nodules ([44]), the tracking of jellyfishes ([45]), and some classification of the sea floor ([46], [47]) or coral reefs [48].

A system that aims to assist the operator in indexing survey video data has been presented in [1]. It automatically performs a reduction of the amount of mission video data to be processed by the scientists. The video material is analyzed through algorithms that detect “significant” changes in context; for instance, the transition from uniform sandy bottom to rocky sea floor, as well as the passing through of big objects such as isolated corals/animals.

The problem of video indexing has recently received much attention for archiving and retrieval in multimedia contexts ([49]–[52]). In most cases, low-level segmentation is facilitated by detecting shot transitions such as cuts or fades, or specific camera motions, which are man-made artifacts of the video stream. The indexing problem addressed here differs significantly in that the typical video material of a mission consists of a continuous video stream that is mainly unedited.

Event detection for automated video surveillance deal with unedited video streams. However, these applications benefit from the availability of *a priori* knowledge [53]. The image background is usually known and stationary, and the interesting events are well defined. Most of such techniques are based on object segmentation and motion analysis.

Instead, for underwater mission video, the camera is mounted on a vehicle, ROV or autonomous underwater vehicles (AUV), and is in permanent motion as the vehicle scans the area to be surveyed, so that the whole image changes continuously. Moreover, the decomposition of the scene into objects does not nec-

Manuscript received April 30, 2002; revised May 8, 2003. This work was part of the EU-MAST project MAS3-CT97-0083, “Advanced ROV Package for Automatic Investigation of Sediments.”

K. Lebart, E. Trucco, and D. M. Lane are with the Oceans Systems Laboratory, School of EPS, Heriot–Watt University, Riccarton Campus, Edinburgh EH14 4AS, U.K. (e-mail: K.Lebart@hw.ac.uk; E.Trucco@hw.ac.uk; D.M.Lane@hw.ac.uk).

C. Smith is with the Department of Marine Ecology and Biodiversity, Institute of Marine Biology of Crete, 71003 Iraklion, Crete, Greece (e-mail: csmith@imbc.gr).

Digital Object Identifier 10.1109/JOE.2003.819314

0364-9059/03$17.00 © 2003 IEEE

essarily yield a relevant representation of content, due to the inherent complexity of the biological and geological structures imaged.

The automatic indexing of video from underwater surveys is, therefore, a very specific and complex problem. Accordingly, the ways to tackle it as well as their evaluation will be specific, as outlined in Section II-A. In the following, we present an assessment protocol for the automatic indexing of underwater survey video data. A set of measures suited to the performance assessment of generic automatic-indexing algorithms is then proposed in Section II-B.

The example algorithm, proposed in [1] (detailed in Section III-B) is used to demonstrate the practical use and benefits of our assessment protocols. The evaluation criteria are used to get an insight into the algorithm behavior, particularly on the influence of the various parameters of the processing. These criteria are then used to benchmark the algorithm with real underwater data, collected during the sea trials of the ARAMIS project and ground truthed by marine scientists. This is presented in Section IV, followed by a discussion of pending issues and future directions of work.

II. Underwater Survey Video Indexing: Performance Assessment

We define the problem of indexing underwater survey video as extracting key frames that sum up the relevant information contained in the video. This information can be application specific: for instance, the presence of trawling marks on the sea floor for fishing impact assessment, a specific animal or plant for population monitoring, or the presence of a man-made object for mine counter-measures applications. In these cases, algorithms can be based on models of these changes or trained using suitable databases. To our knowledge, however, suitable ground-truthed databases (i.e., a collection of real video fully labeled by experts) are not currently available. In their absence, first-generation generic indexing algorithms can be developed, which rely on the detection of changes in the scene imaged. Indeed, a significant change in the scene carries added information that might be relevant to the scientists. One such algorithm has been presented earlier [1] and is used throughout this paper to illustrate the performance-assessment and benchmarking methods presented.

The performance expected from all such indexing algorithms can be expressed and assessed within a similar framework. Next, the specific evaluation requirements for such algorithms are detailed and a set of measures is suggested.

A. Issues for Evaluation

The algorithms need to be evaluated in terms of how well they perform the task of detecting given "events" in a video stream. The two main uses contemplated are

- triggering an alarm in real time when an event of interest occurs during a survey;
- highlighting as much and as concisely possible the global information content of the video record of a mission.

The performance measure must assess

- how many of the actual events are detected ("true positives");
- how many false detections are triggered ("false positives").

We selected two measures that are used in the field of video indexing [49]: *recall* and *precision*.

The sequences to be analyzed contain real images of natural scenes. In this case, the ground truthing of the changes is very subjective, both in the definition of the events and of their moment of occurrence. The spread over time of the relevant events is a general characteristics for underwater survey videos, because of both the nature of what is being observed and the typical vehicle speeds. It is difficult for a human operator trying to ground truth the sequence to decide with a precision of more than a second (20 images) when the change actually occurs.

Section II-B1 describes how the "recall" and "precision" measures are estimated for our algorithm evaluation to take these considerations into account.

Importantly, for the applications and the data considered, the precise time of change is *not* in itself critical: the information retained is the same if an alarm is triggered right at the beginning of a transition or 20 images into the transition. The problem of efficiently extracting the information contained in a survey video tape can actually be cast as retaining as many of the significantly different sea-floor scenes/elements imaged in the video stream, with as little redundancy as possible. To try to quantify this, two complementary measures were introduced: the alarm rates a_s and a_t over, respectively, the stationary or "event-free" chunks and the "transition" or "event-rich" chunks of the video stream. Their estimation is described in Section II-B2.

B. Performance Criteria

1) Recall and Precision: Two indexes can be estimated over test sequences. They are commonly used in the field of information retrieval and video segmentation (e.g., [49])

$$\text{Recall} = \frac{\text{Correct}}{\text{Correct} + \text{Missed}} \tag{1}$$

$$\text{Precision} = \frac{\text{Correct}}{\text{Correct} + \text{False Positive}}. \tag{2}$$

The definitions of "Correct," "Missed," and "False positive" must be specified for the applications considered.

As mentioned in the previous section, it is both subjective and noncritical to try to define the exact position of a transition. Therefore, a margin of 12 images[1] on either side of the labeled transition is taken. A detection is deemed "correct" if it falls within this margin around the ground-truthed transition and "false positive" otherwise. A change has been missed if no change is detected within this margin around a true transition.

2) Sampling Rate: A complementary measure to assess the relevance of the automatic detection is given by the alarm rates a_s and a_t, described next. a_s measures the average alarm rate over the stationary parts of the video, typically a sandy sea floor, and is expected to be low. a_t measures the average alarm rate

[1] This correspond to a 1-s uncertainty if the frame rate is 25 frames/s and should be adjusted depending on the data.

over the "event-rich" parts of the video and is expected to be high.

The effect of the algorithm can be interpreted as a nonuniform sampling process on the video stream at frames deemed significant in terms of added information. These alarm rates correspond to "sampling rates" of the video stream. It is expected that the sampling rate will be significantly higher over transition periods than over stationary zones. These measures highlight the relevance of the algorithm behavior while putting less emphasis on the precise time positioning of the alarms.

To estimate a_s and a_t, we first split the video into sequences in which the content remains mainly the same and sequences in which the content changes relatively rapidly (corresponding to complex transitions between zones). Assume there are N_{zs} sequences of the first type and N_{zt} of the second. We then estimate

$$a_s = \frac{1}{N_{zs}} \sum_{i=1}^{N_{zs}} \frac{N_{a_i}}{N_{f_i}} \tag{3}$$

where N_{zs} is the total number of "event-free" or stationary sequences, N_{a_i} is the number of alarms triggered by the algorithm over the ith stationary sequence, and N_{f_i} is the total number of frames in the same sequence. Also

$$a_t = \frac{1}{N_{zt}} \sum_{i=1}^{N_{zt}} \frac{N_{a_i}}{N_{f_i}} \tag{4}$$

where N_{zt} is the total number of nonstationary (event-rich) sequences, N_{a_i} is the number of alarms triggered by the algorithm over the nonstationary sequence number i, and N_{f_i} is the total number of frames in the nonstationary sequence i.

C. Test Data

Three sets of images were used for the evaluation.

- An outdoor test sequence of 6000 images sampled at 20 frames/s was used for the preliminary study of the algorithm performance. The sequence was shot in a landscaped garden, walking a hand-held camera over different patches of shrubs and floor layers (wood, pebbles, concrete) at about 1-m altitude. The video is, hence, constituted of sequences with natural stationary macro-textures and of the transitions between the different areas. The scene is illuminated by daylight.

  The changes in the sequence have been recorded manually to provide a ground-truth reference for the evaluation and the tuning of parameters. Eight-one "changes" were recorded. An example of a transition in this sequence is represented in Table I.
- Two sequences of real underwater data, ground-truthed by expert scientists, have been used to benchmark the algorithm performance:
  - a sequence of images of sub-sea mission data from trials of the ARAMIS project (example images are shown in Table II);
  - a sequence of underwater images from an archive tape (the "Mixed Dives Tape").

This is detailed in Section IV.

TABLE I
EXAMPLE OF TRANSITION IN THE TEST SEQUENCE

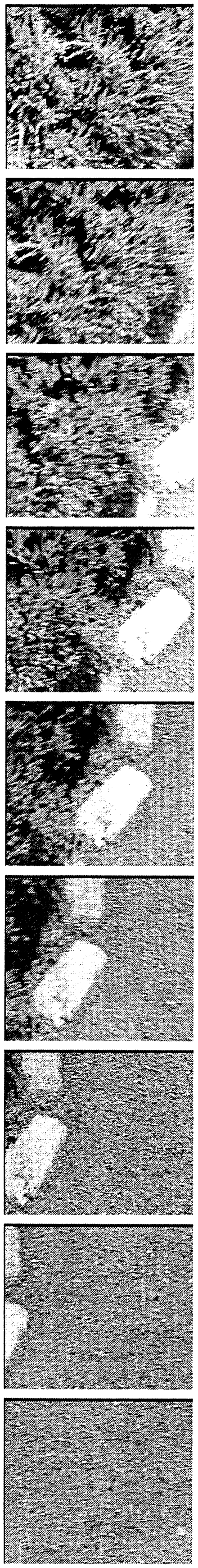
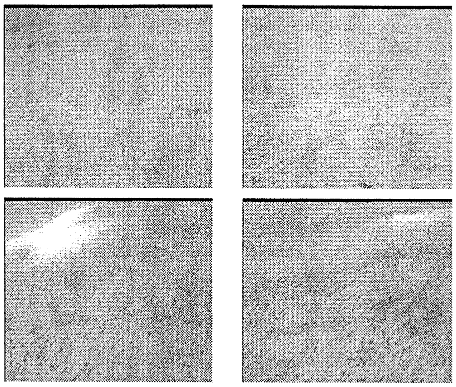


TABLE II
EXAMPLE IMAGES FROM THE UNDERWATER TEST SEQUENCE



III. CHANGE-DETECTION ALGORITHM

An automatic indexing algorithm developed within the European project ARAMIS is presented next. The motivation for the algorithm is outlined in Section III-A. The algorithm principle is then detailed in Section III-B, along with an experimental study of the influence of its parameters.

A. Problem Statement

We consider the case where no prior information on the nature of a survey is available. The working assumption used is that significant changes in the sea floor that is imaged correspond to interesting events.

However, at pixel level, variations in the image can be produced by changes in the real-world scene being imaged, as well as artifacts of the image-acquisition and image-formation processes.

Only the former changes are of interest to us. They must be discriminated from the latter, which for the purpose of our detector are merely noise. In particular, for underwater survey videos the following specific causes of artifacts have to be taken into account:

- *Inter-image variations*
  - Different elements of the scene move over a sequence, changing shape and size as they gradually appear and disappear in the field of view. The motion and changes of objects are changes at pixel level in the image that do not correspond to significant scene changes.
  - The motion of the camera can induce some nonstationary blurring of the images, thereby changing their statistics.
  - The lighting conditions are generally uncontrolled. In shallow water, natural lighting causes rays of lights to sweep through the image, causing strong variations of the gray levels. The propagation of light in water is frequency dependent and the red wavelengths are attenuated very quickly. The same object or zone might loose its red color as the camera moves away.

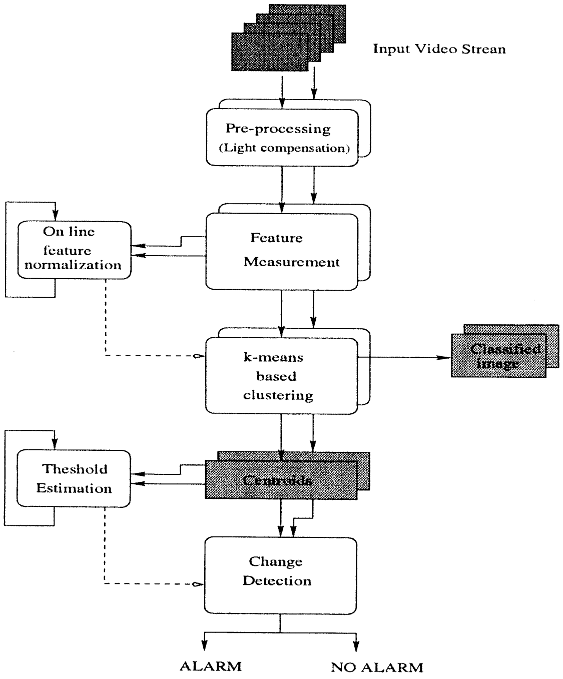


Fig. 1. Change detection from image unsupervised classification.

- *Intra-image disparities*
  - The frequency-dependent attenuation of light causes similar scene elements to look differently, depending on their distance from the camera and light source.
  - Spot lighting induces an unequal illumination pattern on the scene, again creating intensity changes that may not reflect changes of the scene surface.

The algorithm presented next relies on a representation of the image content through classes. The images are analyzed through unsupervised clustering [54]: their content is split into classes, each grouping pixels that have similar characteristics. This provides a representation of the image at a higher level of interpretation than the pixel level. In Section III-B3, we show that this provides robustness toward some of the above-mentioned algorithms.

B. Algorithm Overview

A synopsis of the change detector presented is shown in Fig. 1.

1) Preprocessing: To deal with the nonuniform lighting pattern created by spot lighting, the incoming images are preprocessed through a simple light-compensation algorithm. For this, the spatial pattern of the spot light is first estimated and then used to correct the spatial distribution of lighting. It is assumed that variations in the lighting pattern are much slower than those of the scene. Therefore, a long-term averaging of the images cancels out the short-term details corresponding to the scene. The result is an image I_{illum} of the quasistationary lighting pattern. This image is then used to additively correct the disparities of illumination within the image as follows[2]: let p_{ij} be the pixel

[2] A multiplicative compensation scheme would have more physical grounding, but was found to be numerically unstable. The additive compensation that was implemented has proven to be sufficient to prevent the classification from only identifying lighting differences in the images, rather that differences in the content.

at position i, j of the current image to be compensated; m_{ij} be the pixel at position i, j of I_{illum}; and M be the average gray level value of I_{illum}. Each pixel value is compensated according to

$$p_{ij} := p_{ij} + M - m_{ij}. \quad (5)$$

The result is then truncated so that it lies between 0 and 255.

2) Feature Space Representation of the Image: The incoming image is then split into M nonoverlapping $n \times n$ windows. A set of F statistical features is estimated locally on each such window, yielding a representation of the image content as a cloud of points in an F-dimensional space. It is in this featured space that changes are analyzed.

The features are normalized so that their values lie within the interval [0,1]. This ensures an equal influence of each feature in the clustering. The minimal and maximal value each feature can take is estimated online over the video stream. Unlike local estimation of the extreme values, this enables meaningful cross-image comparison of features values and optimal usage of the feature's dynamic range.

The feature set is a critical factor for the algorithm behavior. Depending on the location, depth, and lighting conditions of the mission, the images display very different characteristics, which can be suitably represented by different features. A case-by-case feature selection would be needed. For the applications and problem considered, a rigorous feature-extraction/selection presents difficulties that can only be overcome by restricting our scope to the detection of well-defined events, such as the occurrence of one specific type of algae in the field of view.

The problem contemplated here is totally unconstrained: we are looking for a wide range of events relating to natural objects in a natural environment, imaged in noncontrolled conditions. A ground-truthed database could only address a limited subset of cases. Moreover, no such database is, to our knowledge, available.

Therefore, here, a feature set comprising three features has been empirically selected.

The set, estimated over 11 × 11-sized windows, comprises

- the mean m_R of the red plane of the image;
- the mean m_B of the blue plane of the image;
- the variance of the gray-level gradient of the image (as a texture descriptor).

This set of features reflects both color and texture properties of the image, which are deemed to characterize marine life in underwater images. Moreover, the texture feature is expected to be relatively insensitive to variations in lighting. It is estimated at low computational cost, which is a requirement for the online implementation of the algorithms.

3) Clustering: Classical methods for change detection in video are based on the local image statistics [49]. These methods are efficient for detecting abrupt changes (typically edit cuts) in a video stream. For our application, as the video stream is mainly unedited, most changes will be smooth transitions. In order to address this major difference, an *unsupervised clustering algorithm (k-means)* [54], [55] is applied to split the cloud of points corresponding to the image in C clusters. Clustering is initialized by the result of the classification of the previous image. This speeds up convergence and guarantees that class labels are consistent across images. The centroids for each class are then calculated. These centroids summarize the statistical information extracted from the image.The measure of change is estimated from the centroid and variance of the entire cloud of points representing the image.

The key clustering parameter is the number of clusters or classes in which to divide up the image points in the feature space. If the number of classes is set to one, there is no clustering and the system is equivalent to histogram-based change detection.

To illustrate how the introduction of clustering can improve the detection of transitions, let us consider an example of what could be a typical change in a mission video: a large rock crosses the field of view within a homogeneous sandy floor. Both techniques rely on a feature representation of the image; therefore, the position of the rock in the image is not in itself important.

If no clustering is done, the cloud of points representing the image in the feature space will drift away from its original position corresponding to "sand only" as more and more points in the image correspond to the rock and their weighting in the cluster characteristics increases. The transition is, therefore, smooth and depends on how many pixels belong to the rock in each frame. The change measure, as a result, might not show any discontinuity. Instead, when the points in the feature space are gathered in multiple clusters based on statistical similarity, then rock pixels are discriminated from sand pixels, provided that the statistical feature set is suitable. Therefore, the algorithm assigns rock and sand components of the image to different classes. This has two consequences.

- As soon as one class is attributed to the rock points, the number of points belonging to the rock only marginally influence the cluster characteristics (centroid and variance) for this class.
- The splitting of the clusters in order to attribute one class to the rock is an event that is localized in time and will appear in the change measure as an instantaneous variation.

We now turn to an experiment validating the assumption that clustering improves the retrieval of transitions. Recall and precision scores of the algorithm have been measured for different numbers of classes (results are displayed in Fig. 2). For a given class number, each curve corresponds to a best-fit curve to the set of (Recall; Precision) values obtained for the parameters S_{km}, α, and W scanning their range of possible values [see (6)–(10)].

For recall values above 0.4, the behavior of the algorithm is similar whatever the number of classes. These situation correspond to high alarm rates, where most of the information is retained by the alarms, along with a rather high number of spurious detections, as reflected by the low precision. For low recall values, however, it can be seen that the precision achieved by the algorithm with 2, 3, and 5 classes is significantly superior to that of the algorithm with no classification stage (one class) or with 10 classes. This means that, for a certain number of changes on the video stream (typically 20% of the relevant changes), the change measure as calculated from the analysis in

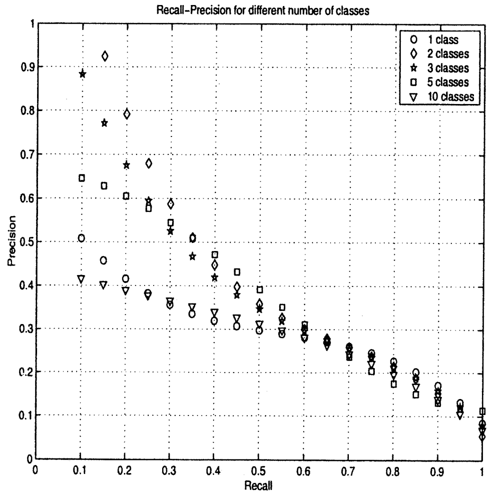


Fig. 2. Influence of the number of classes.

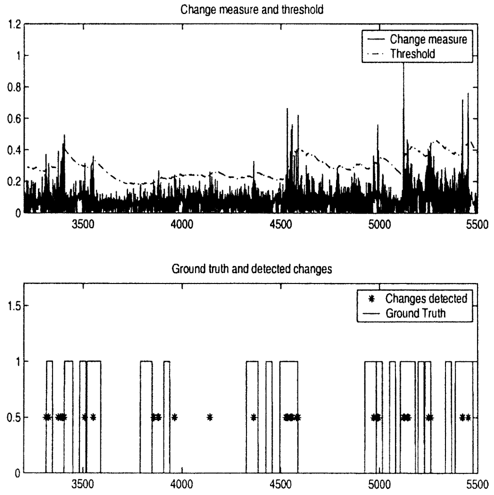


Fig. 3. Change measure and detection.

2, 3, and 5 classes of the image content highlights these changes more reliably than in the two other cases.

In summary, the performance of the algorithm is independent upon the number of classes if high recall is sought. However, significant improvements in precision can be achieved at low recall. Not all changes can be detected by the algorithm; for some, however, the classification stage is of benefit to the performance.

4) Change Detection:

a) Detection: The strategy for change detection is as follows: a variation index $V(n)$ is estimated as a distance between corresponding classes in two successive images. Different indexes have been contemplated, as detailed below (Section III-B4b). The mean $m_V(n)$ of the variation index and its variance $\sigma_V(n)$ are estimated online by a running average, between changes, where $V(n)$ is assumed to be stationary

$$m_V(n+1) = \alpha m_V(n) + (1-\alpha)V(n+1),\ 0 \leq \alpha \leq 1 \quad (6)$$

and

$$\begin{aligned}\mathrm{var}_V(n+1) =& (V(n+1) - m_V(n+1))^2\\ \sigma_V^2(n+1) =& \alpha\sigma_V^2(n) + (1-\alpha)\mathrm{var}_V(n+1).\end{aligned} \quad (7)$$

If a change is detected at time index n, it is likely that the value $V(n+1)$ will be quite high (see Fig. 3). To prevent $V(n+1)$ from biasing the running averages in (6) and (7), its contribution is then attenuated by a factor W

$$m_V(n+1) = \alpha m_V(n) + (1-\alpha).W.V(n+1),\ 0 \leq \alpha \leq 1 \quad (8)$$

and

$$\begin{aligned}\mathrm{var}_V(n+1) =& (V(n+1) - m_V(n+1))^2\\ \sigma_V^2(n+1) =& \alpha\sigma_V^2(n) + (1-\alpha).W.\mathrm{var}_V(n+1).\end{aligned} \quad (9)$$

This is because the change itself should not influence the estimates, which are only meaningful on stationary periods of the signal.

The change test applied is

$$\text{if } V(n) > m_V(n-1) + S_{\mathrm{km}}.\sigma_V(n-1) \Rightarrow \text{Change} \quad (10)$$

where S_{km} is a user-defined parameter that regulates the sensitivity of the detector. The quantity

$$S(n) = m_V(n-1) + S_{\mathrm{km}}.\sigma_V(n-1) \quad (11)$$

is the adaptive threshold above which an alarm is triggered.

b) Measures of Change: Different measures of change have been compared over the test sequence. These measures must capture as many detectable variations as possible, corresponding to the ground-truth variations of the sequence. All the measures were extracted from the classes characteristics; that is, their centroid coordinates and the repartition of their population about these centroids, measured by the variance of the class. The measure selected was the maximum over all classes of the Euclidean distance between the standard deviations, across successive frames.

Let $\sigma_i^f(n)$ be the standard-deviation coordinate for feature f of the ith class in the $n$the frame, N_c be the number of classes, and N_f be the number of features. The change measure $V(n)$ is computed as the max of the Euclidean distance

$$V(n) = \max_{i=1\ldots N_c} \sqrt{\sum_{f=1}^{N_f} (\sigma_i^f(n) - \sigma_i^f(n-1))^2}. \quad (12)$$

It was observed that the "spread" of data clusters, as measured by the standard deviation, is affected more dramatically than the position of the cluster centroids by the relevant changes in the image characteristics. Moreover, the maximum operator enhances the amplitude of the changes when they mostly affect one of the classes. This is expected to correspond to qualitative changes in the image content.

This measure yielded consistently better performance across a wide range of key parameters variations such as number of classes and tuning parameters. The results obtained with this measure also proved to be less sensitive to the variations of the tuning parameters (α, S_{km}). This measure is the only one retained in the following discussion.

Fig. 3 illustrates the behavior of the algorithm over the full sequence for $\alpha = 0.98$, $W = 1.5$, $S_{\mathrm{km}} = 2.9$, and 5 classes. Along the x-axis are the frame numbers in the video sequence.

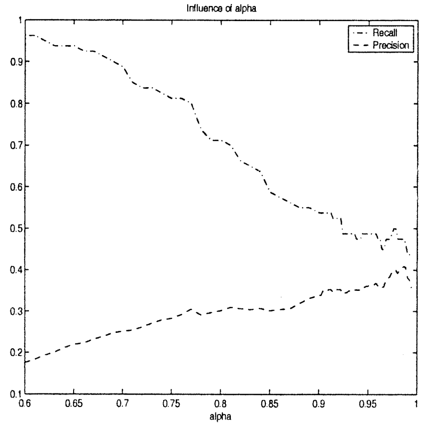


Fig. 4. Influence of running-average parameter α ($W = 1.5, S_{\rm km} = 2.7$).

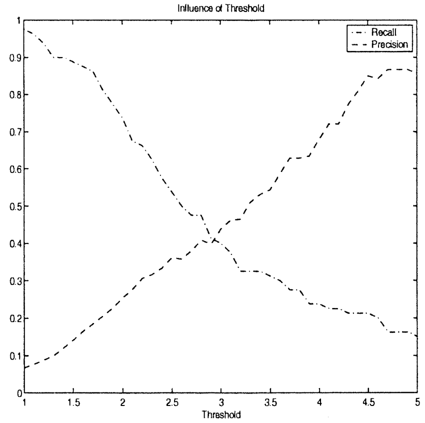


Fig. 5. Influence of threshold factor value $S_{\rm km}$ ($\alpha = 0.9, W = 1.5$).

The continuous line on the top diagram represents the measure of change $V(n)$. The adaptive threshold value $S(n)$ is overlaid (dotted lines). An alarm is triggered when the change signal passes the threshold. In the bottom diagram, the continuous line highlights the stationary zones (value 0) and the nonstationary (or "change rich") zones (value 1). The stars correspond to the alarms. It can be clearly seen that the density of alarms is much higher in nonstationary than in stationary zones.

5) Parameters Influence: Three parameters influence the sensitivity of the detection of changes, namely α and W, the running average parameters defined in (6) and (7) and $S_{\rm km}$, the threshold defined in (10). Their influence is illustrated in terms of compromise between recall and precision.

c) Running Average: Clearly, the choice of α will depend on the typical stationarity length of the video data. Small α put a stronger emphasis on the present, meaning that the threshold will evolve quickly along with the signal. High α will give better estimates of the mean and standard deviation of the process, since the memory of the averaging is then extended, but slower adaptation to changes. Those estimates, however, will be meaningful only if the process remains stationary over the corresponding estimation time. In this respect, W is an *ad hoc* parameter that helps to speed up the convergence of the estimators in (6) and (7) when a nonstationarity (corresponding to a change) is met.

Fig. 4 illustrates the typical influence of α on the algorithm behavior: increasing values of α mainly decrease the number of alarms, therefore increasing the precision and decreasing the recall.

Although both α and W influence the behavior of the algorithm, they were not used as control parameters, since the choice of their value depend mainly on the data for α and on the type of change measure for W. It is to be noted that, within a reasonable range of values, neither of these parameters were critical to the algorithm performance.

For the test sequence (using $V(n)$ as the measure of change), the value for α was set to 0.97 and that of W to 2. However, for underwater surveys where the rate of changes can be expected to be much less, typical suitable values for α are closer to 0.99 for a frame rate of 25 frames/s.

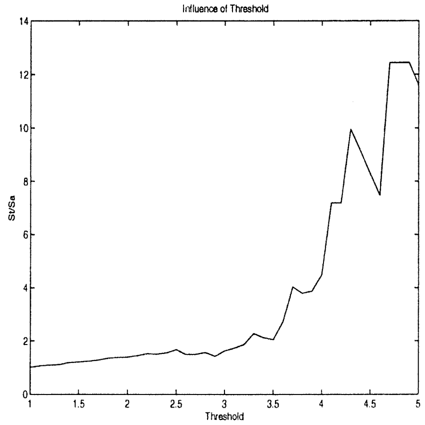


Fig. 6. Influence of threshold on alarm rates ($\alpha = 0.9, W = 1.5$).

d) Threshold: The threshold $S_{\rm km}$, defined in (10), is the key tuning parameter that regulates the compromise between recall and precision performance.

Fig. 5 illustrates the typical influence of $S_{\rm km}$ on the algorithm performance. Decreasing values of $S_{\rm km}$ mainly increases the number of alarms, therefore decreasing the precision and increasing the recall. Fig. 6 represents the evolution of the ratio a_t/a_s where a_s and a_t are the average alarm rates defined in (3) and (4).

As $S_{\rm km}$ increases, the overall number of alarms triggered decreases. Very low $S_{\rm km}$ results in the algorithm triggering very frequent alarms, which are equally divided up among stationary and nonstationary parts of the video stream. This is reflected by the ratio a_t/a_s being close to 1: the sampling of the video is almost uniform. The alarms are, therefore, not highlighting the significant information. However, when $S_{\rm km}$ is increased, the density of alarms over the nonstationary parts increases relative to the number of alarms over the stationary parts ($a_t/a_s > 1$). In other words, even though some relevant changes are missed, the alarms triggered actually highlight the relevant information contained in the video stream: the alarm density is much less over stationary chunks than it is over nonstationary chunks. This is also corroborated by the results displayed in Fig. 2, where high precision is achieved at low recall (top left part of the figure). This means that the most salient changes outlined by the change measure actually correspond to ground-truth changes in the video.

The performance and possible benefits for scientists of the algorithm are assessed next, in a real scientific mission scenario (“ARAMIS trials”) and an archive indexing scenario (“Mixed Dives Tape”). The algorithm is benchmarked against a human expert’s performance.

IV. Test on Underwater Sequences

A. Evaluation Setup and Results

For test purposes, the videos (ARAMIS and “Mixed Dives”) were run through the object-recognition camera system. Although the videos are quite different, the same parameters ($\alpha = 0.97$, $W = 1$, and $S_{\text{km}} = 3.3$) were used for the algorithm. Individual alarms were registered with reference to a frame number. The video section was digitized into approximately 2500 and 3500 images respectively, at a sampling rate of 4 images/s. A marine-biology expert investigated the digital images using the program Thumbnails Plus (running on a PC), which allowed images to be easily loaded into the memory of the computer and to be run in sequence forward and back on an image-by-image basis or as a “video” sequence. The process of analysis consisted of the following.

1) The video was logged by noting any changes (frame number and description) that occurred on a frame-by-frame basis.
2) The video was analyzed by the scientist to note “interesting features” (frame number and description).
3) The algorithm alarms were compared with the “interesting changes” (frame number and description).

B. ARAMIS Trials

As presented here, the algorithm has been integrated within a multisensor underwater survey system developed for the European project “ARAMIS.” The video-processing system consists of an underwater camera fixed on the ARAMIS tool skid and of a surface computer that digitizes the camera signal, processes it online, and outputs analysis results and images at a frame rate of 4 images/s. These can be stored on disk and transmitted through TCP/IP to the central unit of ARAMIS. The system is represented in Fig. 7.

The scientific evaluation of the object-recognition camera was based on a section of video from the ARAMIS scientific mission in the investigation of shallow-water hydrothermal vents from the island of Milos in the Aegean Sea. The test video was taken in Paleohori Bay during one of the programmed missions with the object-recognition camera mounted on the side of the core skid. The camera was looking down and sideways. The video showed a view from a distance of approximately 1 to 10+ m, depending on the altitude of the system, which was mostly stable. The video was collected at approximately 10-m depth in natural lighting conditions on a sunny day.

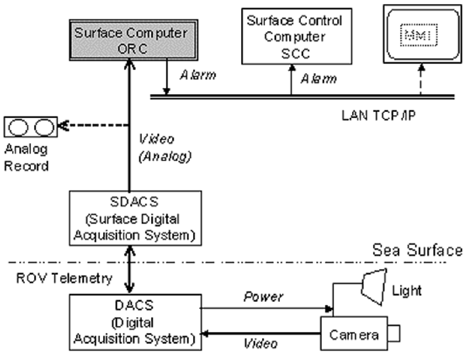


Fig. 7. Aramis object-recognition system.

TABLE III
Correspondence Between the Alarms and the Scientific Features and Between the Alarms. Milos Data

Measure	*Value*
Precision	0.5
Recall	0.35
$\frac{a_t}{a_s}$	3.5

TABLE IV
Correspondence Between the Alarms and the Scientific Features and Between the Alarms. Mixed Dives Tape

Measure	*Value*
Precision	0.57
Recall	0.39

The test video taken during the trial was over a heterogeneous mixed sandy sea bed featuring bare sand, ripples, bioturbated areas, various densities of seagrass, white mats on the sediment surface, and occasional hydrothermal bubble streams rising from the sediment surface. Also present were some larger openings on the sediment surface, probably bioturbation features from burrowing fauna, a fish, and a sponge. Two artifacts were noted on the video: at one point, wave shadows were observed moving across the sediment surface and, at another, the ROV carrier platform turned away from the sun and its diffuse shadow was visible on the sea floor. The system did not produce any alarms corresponding to these two features.

The scientifically interesting features are given in Table V, making a note of the major features in view. From these, “event-free” and “event-rich” zones have been defined. Table X details the list of “event-free” zones used for the calculation of a_t and a_s.

Scientifically interesting views were generally noted as a change in the overall sea-bed conditions with the exception of notation of features of special interest (bubble streams, fish, burrow, and sponge). Some of the features had very sharp borders due to the presence of hydrothermal activity (white patches and surrounding sediment textures). Other biological features (seagrass and bioturbation) generally had increasing or decreasing gradients of density rather than strong borders. Consequently, it was quite difficult to accurately ascertain change to a particular frame number. However, with multiple playbacks a total number of 36 scientifically interesting features were noted (see Table V).

The frame numbers of algorithm alarms are given in Table VII with a frame description provided by the scientific evaluator.

TABLE V
FRAME IDENTIFICATION AND DESCRIPTION OF THE SCIENTIFICALLY INTERESTING FEATURES. MILOS DATA

No.	Frame	Scientific Feature
1	1	Flat sand, sparse seagrass
2	14	Flat sand, sparse seagrass, sponge
3	100	Flat sand, sparse seagrass, dense seagrass patch
4	181	Rippled sand and burrow, denser seagrass
5	310	Flat rippled sand, dark patch
6	362	Rippled sand, dark and white patch
7	400	Rippled sand, dark and white patch, bubble stream
8	412	Flat sand , small burrows
9	466	Flat sand, rippled, white patch and bubble stream
10	516	Bioturbated sand
11	580	Flat sand, white diagonal patch, fish in burrow
12	665	Bioturbation, dense seagrass, white patch
13	717	Flat sand, white/pink patch
14	775	Dark bioturbation patch, white patch
15	803	Flat sand
16	859	Flat sand, dark burrow
17	903	White patch and bioturbated
18	1028	Bioturbated sand, small flat patch
19	1063	Bioturbated sand, white/pink diagonal, small white patches
20	1133	Flat sand, bioturbated patch, ROV shadow
21	1218	White patch, bioturbated sand
22	1248	White patch, flat sand
23	1288	Bioturbated sand
24	1340	Flat sand, sparse seagrass
25	1590	Flat sand, denser seagrass
26	1790	Flat sand, sparse seagrass
27	1922	Flat sand, denser seagrass
28	2030	White patch, rippled sand, very sparse seagrass
29	2102	Flat sand patch, white patch and dense seagrass
30	2167	Bioturbated sand
31	2251	Bioturbated sand, seagrass on low ridge
32	2320	Bioturbated sand
33	2354	White-flat area, bioturbated sand
34	2424	White flat areas, white diagonal, bioturbated, bubble stream
35	2439	Flat sand, white diagonal patch, bubble stream
36	2474	Bioturbated sand, white patch, bubble streams

Table III sums up the values for the performance measures estimated with this data according to the protocol defined in Section II. Approximately 35% of the relevant information is retrieved by the system, with about as many false alarms as there are correct detections (Precision 0.5).

A manual observation by the scientist corroborates these values: in order to have an estimate of the rate of change of sea bed and to compare it with the frequency of the alarms, the image sequence was divided into 500 frame groupings and compared with the alarms in Table IX. This table shows mixed results, which were probably dependent upon the heterogeneity of the environment. With low numbers of changes (frames 1501–2000), the alarm rate was similar to the identified interesting features. At higher rates, there was only a very general similarity, with none of the alarms being consistently precise with the scientific identified features.

It was noted that the alarms, highlighted sequential or close to sequential, frames without obvious "scientific" changes in the video picture. This is an artifact from the change-detection process. Small features, such as a sponge in Frame 14, larger burrow features in the sand, a fish that disappeared into a burrow (Frame 580), and bubble streams could not be detected by the algorithm.

C. Mixed Dives Tape

The second set ("Mixed Dives Tape") of tests is performed on an archive tape. The sequences were filmed in the Aegean Sea during a number of consecutive dives. Sections were edited onto one tape. The video material was acquired by an obliquely forward-looking camera mounted in a free swimming vehicle. Lighting is very uneven, as is the distance from the main plane of view. This is aggravated by the use of a zoom lens. There are 10 sequences:

1) 16.6.95 Iraklion Bay (southern Aegean), 210-m depth: moving forward along a flat muddy sea bed that has a number of diagonal trawl marks of various magnitude from scrapes to deep plough marks;
2) 16.6.95 Iraklion Bay (southern Aegean), 210-m depth: moving forward along a normal flat muddy sea bed with burrows mounds and the sea fan *Leptometra phalangium*;
3) 18.6.95 Psira Island (southern Aegean), 120-m depth: moving up a vertical cliff face with various outcrops and gullies, sponges in the background and large fan-like gorgonian corals;
4) 18.6.95 Psira Island (southern Aegean), 78-m depth: moving upslope above the cliff face on a rocky slope with rocks, gullies, and outcrops;
5) 18.6.95 Psira Island (southern Aegean), 70-m depth: panning camera in close-up position, fixed frame but with rocky motion and fish moving in the fore- and background;
6) 18.6.95 Psira Island (southern Aegean), 70-m depth: fixed frame (with rocky motion), very close closeup with fish moving across the view;
7) 18.6.95 Psira Island (southern Aegean), 62-m depth: fixed view of sponges and with a pan down and zoom out;
8) 26.6.95 Alonisos Island (northern Aegean), 156-m depth: moving forward on muddy sediment stopping at a rock outcrop;
9) 26.6.95 Alonisos Island (northern Aegean), 156-m depth: zoom fixed view of sponge on the top of the rock outcrop;

TABLE VI
FRAME IDENTIFICATION AND DESCRIPTION OF THE SCIENTIFICALLY INTERESTING FEATURES. MIXED DIVES DATA

No.	Frame	Scientific Feature
1	141	Old trawl door mark
2	211	Fresher deeper trawl door mark
3	271	Broken sediment surface
4	493	Start: Muddy sediment normal surface
5	667	Large burrow mound and opening
6	696	Small sediment cloud from moving animal
7	771	Large sediment mound
8	782	Start: Cliff face with corals
9	855	Gorgonian coral mass
10	1085	Gorgonian coral mass
11	1127	Rock gully
12	1190	Lobster (*Palinurus elephas*) antenna under coral mass
13	1272	Gorgonian coral mass
14	1316	Old fishing net on rocks
15	1406	Sponge
16	1460	Gorgonian coral mass
17	1512	Gorgonian coral mass
18	1612	Sponge
19	1664	Gorgonian coral mass
20	1794	Gorgonian coral mass on right edge
21	1830	Sponge
22	1953	Small coral waving in current
23	1999	Gorgonian coral mass
24	2000	Start: Rocky slope above cliff
25	2027	Orange sponges
26	2095	Orange sponges
27	2141	Orange sponges with fish in background
28	2161	Red fishes (*Anthias anthias*) moving
29	2213	Fish shoal on rock
30	2274	Orange sponges
31	2318	Start: Close up of red fishes (*Anthias anthias*)
32	2404	Orange sponges
33	2537	Orange sponges
34	2634	Orange sponges
35	2973	Start: Closer zoom of red fishes (*Anthias anthias*)
36	3057	Start: Trumpet sponge (*Calyx nicaeensis*)
37	3083	Sponges at base of Calyx
38	3137	Start: flat sediment with black ribbon worm (*Bonelia viridis*)
39	3174	Rock with sponges and byrozoans
40	3190	Start: Close up of sponge
41	3281	Start: Close up of octopus (*Octopus vulgaris*)
42	3492	"Zoom out from octopus, black sponge"

TABLE VII
FRAME IDENTIFICATION AND DESCRIPTION OF THE FEATURES IDENTIFIED BY ALARM. MILOS DATA

No.	Frame	Algorithm Alarm Description
1	1	Flat sand, sparse seagrass
2	2	Flat sand, sparse seagrass
3	10	Flat sand, sparse seagrass
4	239	Flat sand, denser seagrass, light bioturbation
5	240	Flat sand, denser seagrass, light bioturbation
6	321	Flat sand sparse seagrass, white patch
7	387	Flat sand, white and dark patch
8	480	Bioturbated, seagrass, white patch, bubble stream
9	481	Bioturbated, seagrass, white patch, bubble stream
10	620	Bioturbated sand, white diagonal, dark patch
11	721	Flat sand, sparse seagrass, white patch border
12	724	Flat sand, sparse seagrass, white patch streak
13	795	White patch border, sea grass band, flat sand
14	796	White patch border, sea grass band, flat sand
15	897	White patch , bioturbated sand
16	913	White patch , bioturbated sand
17	984	White diagonal, bioturbated sand
18	1077	White diagonal, bioturbated sand
19	1107	White diagonal, white patches
20	1123	Diffuse white patch
21	1133	Pink patch, rippled sand, white patch
22	1134	Pink patch, rippled sand, white patch
23	1850	Rippled sand, sparse seagrass
24	1851	Rippled sand, sparse seagrass
25	2017	Seagrass ridge, white patch
26	2072	Denser seagrass, white patch
27	2083	Denser seagrass, white patch
28	2197	Bioturbated sand, seagrass ridge
29	2344	Flat patch, bioturbated sand
30	2418	Flat patch bioturbation border, flat patch diagonal, bubble stream
31	2463	Bioturbation, white diagonal, bubble stream

10) 26.6.95 Alonisos Island (northern Aegean), 156-m depth: zoom fixed view of an octopus in pebble field.

Table VI gives the scientifically interesting features, as identified by an expert biologist, as well as the frame numbers for cuts, with a note of the major features in view. In total, 42 "interesting events" were noted. In this specific case of an archive tape made of relatively short and widely varied clips, the interesting events are spread evenly across the tape. As there are no long sequences with no events, the measure a_t/a_s cannot be meaningfully estimated.

The frame numbers of algorithm alarms are given in Table VIII, with a frame description provided by the scientific evaluator. The total number of good detections is 17, 9 of which are cuts. There are 11 false alarms and 25 misses.

Table IV sums up the values for the performance measures estimated with this data according to the protocol defined in

TABLE VIII
FRAME IDENTIFICATION AND DESCRIPTION OF THE FEATURES IDENTIFIED BY ALARM. MIXED DIVES DATA

Number	**Frame**	Description
1	**125**	Scrape marks and old trawl door mark
2	**126**	Scrape marks and old trawl door mark
3	**214**	Fresh deep trawl door mark
4	**370**	Smooth relatively unmarked sediment
5	**438**	Flat sediment with old scrape mark
6	**493**	Start: Bioturbated muddy sediment burrows and *Leptometra phalangium*
7	**648**	Bioturbated sediment with *Leptometra phalangium*;Frame jump in sequence
8	**771**	Mound and burrow on sediment
9	**782**	Start: Cliff face with corals
10	**1124**	Rock gully
11	**1359**	Old fishing net on rocks
12	**1393**	Sponges
13	**1470**	"Gorgonian coral mass, sponge mass on base"
14	**1663**	Gorgonian coral mass
15	**1666**	Gorgonian coral mass
16	**1745**	Rock outcrop
17	**1899**	Sand shelf
18	**1901**	Sand shelf
19	**2000**	Start: Rocky slope above cliff
20	**2141**	Orange sponges with red fish (*Anthias anthias*)
21	**2318**	Start: Close up of red fishes (*Anthias anthias*)
22	**2718**	Orange sponges
23	**2877**	Orange sponges
24	**2973**	Start: Closer zoom of red fishes (*Anthias anthias*)
25	**3057**	Start: Trumpet sponge (*Calyx nicaeensis*
26	**3137**	Start: flat sediment with black ribbon worm (*Bonelia viridis*)
27	**3190**	Start: Close up of sponge
28	**3281**	Start: Close up of octapus (*Octopus vulgaris*)
29	**3355**	Octopus out of focus Frame jump out of focus
30	**3357**	Octopus Frame jump
31	**3358**	Octopus Frame jump

TABLE IX
RATE OF ALARMS AGAINST THE RATE OF SCIENTIFICALLY OBSERVED INTERESTING OBJECTS

Frame	Science	Alarm
0-500	9	9
501-1000	8	8
1001-1500	7	5
1501-2000	3	2
2001-2500	9	7
Total	36	31

TABLE X
EVENT-FREE ZONES ON THE MILOS VIDEO DATA

Start frame	End frame
27	76
111	153
191	0282
591	0637
917	1001
1359	1569
1604	1769
1804	1898
1933	2008
2262	2362
2401	2304
Apx Total Frames	1138

Section II. Approximately 40% of the relevant information is retrieved by the system, with about as many false alarms as there are correct detections. It is to be noted that, in this specific test, performance is boosted by the cuts. These are very obvious changes and all are detected correctly. Note that the recall and precision fall to 0.24 and 0.42, respectively, if cuts are not considered.

One reason for this is that the content of edited archive tapes is much richer than that of a typical mission. The type of events noted by the scientist are very specific (particular species of fauna and flaura, etc.). The algorithm is looking for significant changes, which in these cases rarely coincide with the events. There is, therefore, an increased semantic gap between the scientist's interests and what the algorithm can detect and, hence, a stronger mismatch between what triggers the alarms and what interests the scientists.

Another aspect to be considered is the fact that the sequences, now very different in content, are relatively short. The adaptive threshold of the algorithm, as well as the online feature-normalization mechanism, require relatively long stationary sequences to be effective. On fast-varying data such as the Mixed Dives tape, the detection threshold is likely to be mistuned, resulting in the detection of only the most salient changes.

D. Discussion

The camera alarms were triggered by image events defined precisely within the algorithm discussed above. In comparison, the scientist can be both subjective and intuitive; the accuracy is highly dependent upon experience and what exactly is being investigated (general view, specific objects, etc.). The success rate of the alarms in terms of frame-by-frame comparison (+/− 12 frames) with the scientifically identified features was given at 30-35%. This may not seem very accurate,[3] but must be put in context. First, the task tackled by the algorithm is extremely complex, all the more so in the absence of supervision in the feature-estimation and classification steps. We also need to consider what the two systems (computer and scientist) were looking for. The scientist was looking for scientifically interesting features. This ranged from ill-defined details of special interest (bubble stream, sponge, fish, and burrow), which make

[3] As a comparison, if 31 alarms were triggered randomly among the 25 000 frames, the chance of 15 falling within the 775 images defined around the "alarms frames" as being correct alarms is only 2%.

up a very small part of the screen (less than 5%), to substantial background changes (flat and bioturbated sand, white patch, and seagrass). With certain features (notably white patches), there were very strong borders that could be precisely observed. Generally speaking, with biological features (bioturbation and seagrass), there tends to be an incremental change with potentially little noticeable difference between sequences of hundreds of frames (cf. frames 1501–2000 in the Milos data). If the small features of special interest are put to one side, the scientist was looking for complete changes in view, i.e., from a seagrass bed to a large white patch. The object-recognition system seemed to be better at picking up the borders of change between the seagrass and the white patch, whereas the scientist would note the change beyond the border unless the border was a consistent item.

The object-recognition camera system can run in real time (with a frame rate of 5 images/s), or in post-processing mode. The scientist is not fast enough in most cases to analyze the video in real time, although some type of notes can be made for post-processing referral. For the scientific comparison and identification of interesting features, the scientist had to post-process the video on more or less a frame-by-frame basis, playing the sequence through a number of different times and taking a considerable amount of time. The identification by the system of the borders allows the scientist to highlight areas of interest for post-processing. The capability of detecting cuts in the Mixed Dives data, as well as interesting events, also demonstrate the potential of the algorithm as an archive-indexing tool. This would allow the system to be used for screening and summarizing video material either online or in post-processing mode.

V. Conclusion and Future Work

We have outlined a performance-evaluation and benchmarking framework for underwater survey video-indexing purposes and have illustrated how it could be used to tune, study, and evaluate the performance of underwater video-indexing algorithms through one example algorithm.

The algorithm used for illustration addresses the very challenging problem of finding events in an underwater video that are of interest to marine scientists, either biologists or geologists. The imaging conditions are not controlled in any way, which induces a high variability in the images. Moreover, the scenes observed are natural and, therefore, highly variable as well. The somewhat subjective definition of what is an “interesting” event for the scientists adds to the challenge of this problem. The method relies on a unsupervised, high-level analysis of the image content based on the local statistical properties of the images. The system currently proceeds to nonuniform sampling of the video stream. This allows us to retrieve some, but not all, of the interesting information with a considerable reduction of redundancy. In the tests carried out on an underwater sequence, about 30% of the desired information was retained and highlighted through only 31 and 28 images, respectively, of the 2500 and 3500 frames constituting the original sequences. The main tuning parameter overall is the threshold S_{km}, which regulates the compromise between recall and precision. The robustness of the approach enables one to use similar sets of parameters for widely differing video material, such as the in-air test sequence and the underwater survey sequences used in this paper.

This work constitutes a first building block toward the automated indexing of nonedited video-survey material. For the algorithm, the limited performance obtained is a direct reflection of the extremely small amount of *a priori* knowledge available. However, the algorithm is a starting point structure from which more specific detection problems can be addressed. Two stages in the current algorithm are currently unsupervised: the feature selection and the classification. It is expected that much can be gained by introducing some training and supervision at both these stages. This, in turn, will narrow down the purpose of the algorithm. When the purpose of the video survey is to monitor a known area of sea bed, data previously collected can be used to train the algorithm, so that the different types of sea bed known to be present can be reliably identified. These issued are currently being explored within the European project AMASON (EVK3-CT-2001-00 059), with application to trawling impact assessment and coral monitoring. The performance-evaluation protocol presented here is meant to assess and benchmark these future algorithms.

Acknowledgment

The authors would like to thank A. Grehan of the University of Galway, Ireland, for many useful discussions.

References

[1] K. Lebart, E. Trucco, and D. M. Lane, “Real-time automatic sea-floor change detection from video,” in *Proc. MTS/IEEE OCEANS ’00*, Providence, RI, Sept. 2000, pp. 337–343.

[2] I. R. MacDonald and S. K. Juniper, “Sipping form the firehose: How to tame the data flow from ROV operations,” *MTS J.*, vol. 31, no. 3, pp. 61–67, 1997.

[3] R. L. Marks, H. H. Wang, M. J. Lee, and S. M. Rock, “Automatic visual station keeping of an underwater robot,” in *Proc. MTS/IEEE OCEANS ’94*, vol. 2, Brest, France, 1994, pp. 137–142.

[4] L. Jin, X. Xu, S. Negahdaripour, C. Tsukamoto, and J. Yuh, “A real-time vision-based stationkeeping system for underwater robotics applications,” in *Proc. MTS/IEEE OCEANS ’96*, vol. 3, Ft. Lauderdale, FL, 1996, pp. 1076–1081.

[5] K. N. Leabourne, S. M. Rock, S. D. Fleischer, and R. Burton, “Station keeping of an ROV using vision technology,” in *Proc. MTS/IEEE OCEANS ’97*, vol. 1, Halifax, NS, Canada, 1997, pp. 634–640.

[6] X. Xu and S. Negahdaripour, “Automatic optical station keeping and navigation of an R.O.V.; sea trial experiments,” in *Proc. MTS/IEEE OCEANS ’99*, vol. 1, Seattle, WA, 1999, pp. 71–76.

[7] J. O. Hallset, “Simple vision tracking of pipelines for an autonomous underwater vehicle,” in *IEEE Int. Conf. Robotics and Automation*, vol. 3, Sacramento, CA, 1991, pp. 2767–2772.

[8] P. Rives and J. J. Borrelly, “Underwater pipe inspection task using visual serving techniques,” in *IEEE Int. Conf. Intelligent Robots and Systems*, vol. 1, Grenoble, France, 1997, pp. 63–68.

[9] B. A. A. P. Balasuriya, M. Takai, W. C. Lam, T. Ura, and Y. Kuroda, “Vision based autonomous underwater vehicle navigation: Underwater cable tracking,” in *IEEE Oceans Conf. Rec.*, vol. 2, 1997, pp. 1418–1424.

[10] J. Kojima, Y. Kato, K. Asakawa, and N. Kato, “Experimental results of autonomous underwater vehicle “aqua explorer 2” for inspection of underwater cables,” in *IEEE Oceans Conf. Rec.*, vol. 1, 1998, pp. 113–117.

[11] J.-F. Lots, D. M. Lane, and E. Trucco, “Application of 2 1/2 d visual servoing to underwater vehicle station-keeping,” in *MTS/IEEE OCEANS ’00*, Providence, RI, Sept. 2000, pp. 1257–1264.

[12] R. L. Marks, S. M. Rock, and M. J. Lee, “Real-time video mosaicking of the ocean floor,” in *Proc. 1994 Symp. Autonomous Underwater Vehicle Technology*, Cambridge, MA, 1994, pp. 21–27.

[13] S. D. Fleisher, R. L. Marks, S. M. Rock, and M. J. Lee, "Improved real-time video mosaicking of the ocean floor," in *Proc. MTS/IEEE OCEANS '95* , vol. 1, San Diego, CA, 1995.

[14] S. D. Fleischer, H. H. Wang, S. M. Rock, and M. J. Lee, "Video mosaicking along arbitrary vehicle paths," in *Proc. 1996 Symp. Autonomous Underwater Vehicle Technology*, Monterey, CA, 1996, pp. 293–299.

[15] N. Gracias and J. Santos-Victor, "Automatic mosaic creation of the ocean floor," in *IEEE Oceans Conf. Rec.*, vol. 1, 1998, pp. 257–262.

[16] J. Guo, S. W. Cheng, and J. Y. Yinn, "Underwater image mosaicking using maximum a posteriori image registration," in *Proc. 2000 Int. Symp. Underwater Technology.*, Tokyo, Japan, May 23–26, 2000, pp. 393–398.

[17] A. Khamene and S. Negahdaripour, "Building 3-d elevation maps of sea-floor scenes from underwater stereo images," in *Proc. MTS/IEEE OCEANS '99*, vol. 1, Seattle, WA, 1999, pp. 64–70.

[18] H. Singh, J. Howland, and D. Yoerger, "Quantitative photomosaicking of underwater imagery," in *Proc. MTS/IEEE OCEANS '98*, vol. 1, Nice, France, 1998.

[19] D. Wettergreen, C. Gaskett, and A. Zelinsky, "Development of a visually-guided autonomous underwater vehicle," in *Proc. MTS/IEEE OCEANS '98*, vol. 1, Nice, France, 1998.

[20] S. Negadaripour, X. Xu, A. Khamene, and Z. Awan, "3-D motion and depth estimation from sea-floor images for mosaic-based station-keeping and navigation of rovs/auvs and high-resolution sea-floor mapping," in *Proc. 1998 Workshop on Autonomous Underwater Vehicles*, Cambridge, MA, 1998, pp. 191–200.

[21] A. Huster, S. D. Fleischer, and S. M. Rock, "Demonstration of a vision-based dead-reckoning system for navigation of an underwater vehicle," in *Proc. IEEE Symp. Autonomous Underwater Vehicle Technology*, Cambridge, MA, 1998, pp. 185–189.

[22] S. D. Fleischer, S. M. Rock, and R. Burton, "Global position determination and vehicle path estimation from a vision sensor for real-time video mosaicking and navigation," in *IEEE Oceans Conf. Rec.*, vol. 1, Halifax, NS, Canada, 1997, pp. 641–647.

[23] R. Garcia, J. Batlle, X. Cufi, and J. Amat, "Positioning an underwater vehicle through image mosaicking," in *IEEE Int. Conf. Robotics and Automation*, vol. 3, Seoul, Korea, 2001, pp. 2779–2784.

[24] X. Xu and S. Negahdaripour, "Application of extended covariance intersection principle for mosaic-based optical positioning and navigation of underwater vehicles," in *IEEE Int. Conf. Robotics and Automation*, vol. 3, Seoul, Korea, 2001, pp. 2759–2766.

[25] S. Negahdaripour, X. Xu, and L. Jin, "Direct estimation of motion from sea floor images for automatic station-keeping of submersible platforms," *IEEE J. Oceanic Eng.*, vol. 24, pp. 370–382, July 1999.

[26] S. Negahdaripour, C. H. Yu, and A. H. Shokrollahi, "Recovering shape and motion from undersea images," *IEEE J. Oceanic Eng.*, vol. 15, pp. 189–198, July 1990.

[27] S. Negahdaripour, "Revised definition of optical flow: Integration of radiometric and geometric cues for dynamic scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 961–979, Sept. 1998.

[28] F. Spindler and P. Bouthemy, "Real-time estimation of dominant motion," in *IEEE Int. Conf. Robotics and Automation*, Leuven, Belgium, 1998, pp. 1063–1068.

[29] C. Plakas and E. Trucco, "Developing a real-time, robust, video, tracker," in *MTS/IEEE OCEANS '00*, Providence, RI, 2000, pp. 1345–1352.

[30] A. Branca, E. Stella, and A. Distante, "Autonomous navigation of underwater vehicles," in *Proc. MTS/IEEE OCEANS '98*, Nice , France, 1998.

[31] T. Tommasini, A. Fusiello, V. Roberto, and E. Trucco, "Robust feature tracking in underwater video sequences," in *Proc. MTS/IEEE OCEANS '98* , Nice , France, 1998.

[32] F. Aguirre, J. M. Boucher, and J. J. Jacq, "Underwater navigation by video sequence analysis," in *Proc. 10th Int. Conf. Pattern Recognition*, vol. ii, Atlantic City, NJ, 1990, pp. 537–539.

[33] S. M. Zanoli and P. Zingaretti, "Underwater imaging system to support ROV guidance," in *IEEE Oceans Conf. Rec.*, vol. 11, 1998, pp. 257–268.

[34] P. Zingaretti and S. M. Zanoli, "Robust real-time detection of an underwater pipeline," *Eng. Applicat. Artificial Intell.*, vol. 11, no. 2, pp. 257–268, 1998.

[35] A. Grau, J. Climent, and J. Aranda, "Real-time architecture for cable tracking using texture descriptors," in *IEEE Oceans Conf. Rec.*, vol. 3, Nice, France, 1998.

[36] A. J. R. Fairweather, A. R. Greig, and M. Hodgetts, "Robust scene interpretation of underwater image sequences," in *Proc. 1997 6th Int. Conf. Image Processing and Its Applications.*, vol. 442/443, Dublin, Ireland, 1997, pp. 660–664.

[37] S. Matsumoto and Y. Ito, "Real-time vision-based tracking of submarine cables for AUV/ROV," in *Proc. MTS/IEEE OCEANS '95 Conf.*, vol. 3, San Diego, CA, 1995, pp. 1997–2002.

[38] A. Ortiz, G. Olivier, and J. Frau, "A vision system for underwater real-time control tasks," *Proc. MTS/IEEE OCEANS '97 Conf.*, vol. 2, pp. 1425–1430, Oct. 6–9, 1997.

[39] G. L. Foresti, S. Gentili, and M. Zampato, "Vision-based system for autonomous underwater vehicle navigation," in *IEEE Oceans Conf. Rec.*, vol. 1, 1998, pp. 195–199.

[40] J. Kojima, Y. Kato, and K. Asakawa, "Autonomous underwater vehicle for inspection of submarine cables," in *Proc. Int. Offshore and Polar Engineering Conf.*, vol. 2, Brest, France, 1999, pp. 458–462.

[41] Y. Petillot, K. Lebart, A. Cormack, and D. Lane, "Seetrack, a system for post mission analysis of AUVdata products," in *Proc. GOATS '00 Conf.* La Spezia, Italy, 2001.

[42] D. M. Kocak, N. da Vitoria Lobo, and E. Widder, "Computer vision techniques for quantifying, tracking, and identifying bioluminescent plankton," *IEEE J. Oceanic Eng.*, vol. 24, pp. 81–95, Jan. 1999.

[43] R. Li, H. Li, W. Zou, R. G. Smith, and T. A. Curran, "Quantitative photogrammetric analysis of digital underwater video imagery," *IEEE J. Oceanic Eng.*, vol. 22, pp. 364–375, Apr. 1997.

[44] C.-Y. Park, S.-H. Park, C.-W. Kim, J.-. K. Kang, and K.-H. Kim, "Image analysis technique for exploration of manganese nodules," *Marine Georesources and Geotechnology*, vol. 17, pp. 371–386, 1999.

[45] J. Rife and S. M. Rock, "Visual tracking of jellyfish in situ," in *Proc. Int. Conf. Image Processing*, vol. 1, Thessaloniki, Greece, 2001, pp. 289–292.

[46] A. Olmos, E. Trucco, K. Lebart, and D. M. Lane, "Detecting ripple patterns in mission videos," in *Proc. MTS/IEEE OCEANS '00*, Providence, RI, Sept. 2000, pp. 331–335.

[47] N. Pican, E. Trucco, M. Ross, D. M. Lane, Y. Petillot, and I. Tena Ruiz, "Texture analysis for seabed classification: Co-occurrence matrices vs. self-organizing map," in *IEEE/OES OCEANS '98 Conf.*, Nice, France, Sept. 1998.

[48] M. Soriano, S. Marcos, C. Salorna, M. Quibilan, and P. Alino, "Image classification of coral reef components from underwater color video," in *Proc. MTS/IEEE OCEANS '01 Conf.*, vol. 2, 2001, pp. 1008–1013.

[49] R. Brunelli, O. Mich, and C. M. Modena, "A survey on the automatic indexing of video data," *J. Visual Commun. Image Representat.*, vol. 10, pp. 78–112, 1999.

[50] J. S. Boreczky and L. A. Rowe, "Comparison of video shot boundary detection techniques," in *Proc. SPIE, Storage and Retrieval for Still Image and Video Databases IV*, vol. 2670, San Jose, CA, 1996, pp. 170–179.

[51] H. H. Yu and W. Wolf, "A hierarchical multiresolution video shot transition detection scheme," *Comput. Vision Image Understanding*, vol. 75, no. 1/2, pp. 196–213, 1999.

[52] B. Gunsel, A. M. Ferman, and A. M. Tekalp, "Temporal video segmentation using unsupervised clustering and semantic object tracking," *J. Electron. Imaging*, vol. 7, no. 3, pp. 592–604, 1998.

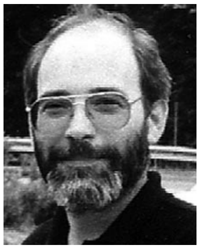
[53] F. Bremond and M. Thonnat, "Issues of representing context illustrated by video-surveillance applications," *Int. J. Human-Computer Stud.*, vol. 48, pp. 375–391, 1998.

[54] A. K. Jain, R. P. W. Duin, and J. Ma, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, 2000.

[55] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. New York: Academic, 1998.

Katia Lebart received the Ph.D. degree in signal processing jointly from the University of Sussex, Sussex, U.K., and the University of Rennes I, Rennes, France, in 1999.

She is a Lecturer at Heriot-Watt University, Edinburgh, U.K. Her research interests include applications of statistical signal processing with a recent focus on image processing and video analysis applied to underwater video images.



Chris Smith is a Benthic Ecologist with the Insitute of Marine Biology of Crete, Greece, since 1988. He is involved in investigations of soft sedimentary ecosystems with an emphasis on the application of modern technologies (towed video, side-scan sonar, ROV, and sediment profile imagery). His research work is funded through collaborative European Union projects.

Emanuele Trucco received the Ph.D. degree in electronic engineering from the University of Genoa, Genoa, Italy.

He has 16 years of experience in computer vision research and applications to subsea robotics, aquaculture automation, industrial inspection, and telepresence. He is a Senior Lecturer at Heriot-Watt University, Edinburgh, U.K.

David M. Lane is Professor in the School of Engineering and Physical Sciences, Heriot Watt University, Edinburgh, U.K., and Director of the university’s Ocean Systems Laboratory. He is also cofounder of SeeByte Ltd., Edinburgh, U.K. He has published over 100 journal and conference papers on tethered and autonomous underwater vehicles, subsea robotics, image processing, and advanced control.