Journal of Applied Ecology 2005
42, 999–1011

METHODOLOGICAL INSIGHTS

An astronomical pattern-matching algorithm for computer-aided identification of whale sharks *Rhincodon typus*

Z. ARZOUMANIAN,* J. HOLMBERG† and B. NORMAN‡

**Universities Space Research Association, NASA Goddard Space Flight Center, Greenbelt, MD 20770, USA; †3433 NE 44th Avenue, Portland, Oregon, 92713, USA; and ‡ECOCEAN, c/o Centre for Fish and Fisheries Research, Murdoch University, South St., Murdoch, WA 6150, Australia*

Summary

1. The formulation of conservation policy relies heavily on demographic, biological and ecological knowledge that is often elusive for threatened species. Essential estimates of abundance, survival and life-history parameters are accessible through mark and recapture studies given a sufficiently large sample. Photographic identification of individuals is an established mark and recapture technique, but its full potential has rarely been exploited because of the unmanageable task of making visual identifications in large data sets.
2. We describe a novel technique for identifying individual whale sharks *Rhincodon typus* through numerical pattern matching of their natural surface 'spot' colourations. Together with scarring and other markers, spot patterns captured in photographs of whale shark flanks have been used, in the past, to make identifications by eye. We have automated this process by adapting a computer algorithm originally developed in astronomy for the comparison of star patterns in images of the night sky.
3. In tests using a set of previously identified shark images, our method correctly matched pairs exhibiting the same pattern in more than 90% of cases. From a larger library of previously unidentified images, it has to date produced more than 100 new matches. Our technique is robust in that the incidence of false positives is low, while failure to match images of the same shark is predominantly attributable to foreshortening in photographs obtained at oblique angles of more than 30°.
4. We describe our implementation of the pattern-matching algorithm, estimates of its efficacy, its incorporation into the new ECOCEAN Whale Shark Photo-identification Library, and prospects for its further refinement. We also comment on the biological and conservation implications of the capability of identifying individual sharks across wide geographical and temporal spans.
5. *Synthesis and applications.* An automated photo-identification technique has been developed that allows for efficient 'virtual tagging' of spotted animals. The pattern-matching software has been implemented within a Web-based library created for the management of generic encounter photographs and derived data. The combined capabilities have demonstrated the reliability of whale shark spot patterns for long-term identifications, and promise new ecological insights. Extension of the technique to other species is anticipated, with attendant benefits to management and conservation through improved understanding of life histories, population trends and migration routes, as well as ecological factors such as exploitation impact and the effectiveness of wildlife reserves.

Key-words: conservation, marine and fisheries management, mark–recapture, photographic identification, population studies

Journal of Applied Ecology (2005) **42**, 999–1011
doi: 10.1111/j.1365-2664.2005.01117.x

Correspondence: Z. Arzoumanian, Universities Space Research Association, NASA Goddard Space Flight Center, Greenbelt, MD 20770, USA (fax +01 301 286 1684; e-mail zaven@milkyway.gsfc.nasa.gov).

Introduction

The whale shark *Rhincodon typus* Smith 1829 (Melville 1984) is the world's largest fish species but is both rare and poorly studied. One of approximately 370 shark species (Last & Stevens 1994), it is a member of the order Orectolobiformes, predominantly bottom-dwellers such as the wobbegong and carpet sharks (e.g. *Orectolobus ornatus* and *Hemiscyllium ocellatum*, respectively; Compagno 1988). Whale sharks have a broad distribution in tropical and warm temperate seas, usually between latitudes 30°N and 35°S (Last & Stevens 1994; Norman 1999). The World Conservation Union (IUCN) *Red List of Threatened Species* (Baillie, Hilton-Taylor & Stuart 2004) lists the whale shark as vulnerable to extinction, as a result of directed fisheries, high value in international trade, a highly migratory nature, a *K*-selected life history and generally low abundance (Norman 2000).

As with any exploited species, effective management and conservation practices for whale sharks are best derived from a sound ecological foundation (Ormerod 2003). A number of outstanding questions in whale shark ecology may be addressed through collection and subsequent collation of sighting data (Norman 2004): mark and recapture studies are possible whenever animals can be 'marked', or otherwise identified, and 'recaptured', or identified later by resighting (Lettink & Armstrong 2003). Analysis of the resulting data can be used to estimate abundance, survival, recruitment and population growth rates over time (Thompson, White & Gowan 1998). Importantly, such research can provide improved assessments of the global conservation status of a species. Conventional tagging of whale sharks, however, has met with limited success.

Whale sharks are born with unique body pigmentation that is retained throughout their lives (Norman 2004). This natural patterning of lines and spots shows no evidence of significant change over years and may therefore be used to identify individual sharks (Taylor 1994; Norman 1999): its uniqueness has been corroborated by traditional tagging and identifications made based on scarring and other visual markers. Through the combination of photographed encounters and spot-pattern matching, a shark may be 'tagged' without physical contact or interference with the animal. In an early effort, Norman (1999) established a photo-identification library of whale sharks at Ningaloo Reef, Western Australia, with photographs of individual sharks examined by eye for identifying characteristics, including spot patterns.

Photo-identification catalogues for some species, notably marine mammals, have been in use for two decades or more (Hammond, Mizroch & Donovan 1990), with most relying on visual matching of individuals. While it may be possible to manage small numbers of photographs and identify individuals by eye, the process becomes inefficient and unreliable when collating data from many animals sighted in many regions around the world. The availability of large data sets has rendered manual photo-identification unfeasible, motivating the development of computer-aided techniques for scanning photographic catalogues accurately and efficiently. Recent computer-aided efforts for marine mammals have focused on the characteristic shapes and colouring of fins and flukes; these include EUROPHLUKES (Evans 2003), DARWIN (Wilkin, Debure & Roberts 1998), The Dolphin Project (Lapolla 2005) and the Mid-Atlantic Bottlenose Dolphin Catalogue (Urian 2005). The last two use the Finscan software (Hillman *et al*. 2003) to identify individuals by the shapes of their dorsal fins.

In this paper, we present a numerical method for identifying individual whale sharks by the unique patterning of their surface spots. Our technique is adapted from an algorithm developed within the astronomical community for stellar pattern recognition. It has been incorporated into the ECOCEAN Whale Shark Photo-identification Library (http://photoid.whaleshark.org), an online database facility that archives digital images submitted by researchers and other interested parties. With a sophisticated pattern-matching capability and a growing library of images, we anticipate that a scientifically valuable number of individual sharks will be identified across wide geographical and temporal spans, improving our understanding of whale shark life histories, migration patterns and demographics. The technologies upon which both the Library and the pattern-matching method are based are generic and can be applied in principle to any species that exhibits distinctive skin patterning. The resulting ecological insights should helpfully inform conservation and management efforts.

Materials and methods

PROVENANCE OF PHOTO-IDENTIFICATION LIBRARY DATA

To support the collection and centralization of biological data by wildlife researchers, the Shepherd Project was begun in 2002 with the goal of creating a reusable World Wide Web-based catalogue framework for the management of mark–recapture data accumulated by a global research community, ecotourists and government agencies. This framework combines an object-orientated database, image management and protection functionality, an extensible programming interface and parameter search capability. A data export facility to support trending and population analyses using Microsoft Excel or Program Mark is also included. The Shepherd Project effort was completed in 2004 and first employed in the ECOCEAN Whale Shark Photo-identification Library. The Library, built upon a J2EE software platform (Sun Microsystems Inc., Santa Clara, California, USA), is a repository for whale shark spot-pattern data and the photographs from which they are derived. Basic information required to accompany photographs includes (i) sighting date and location, (ii) sex and size of the animal

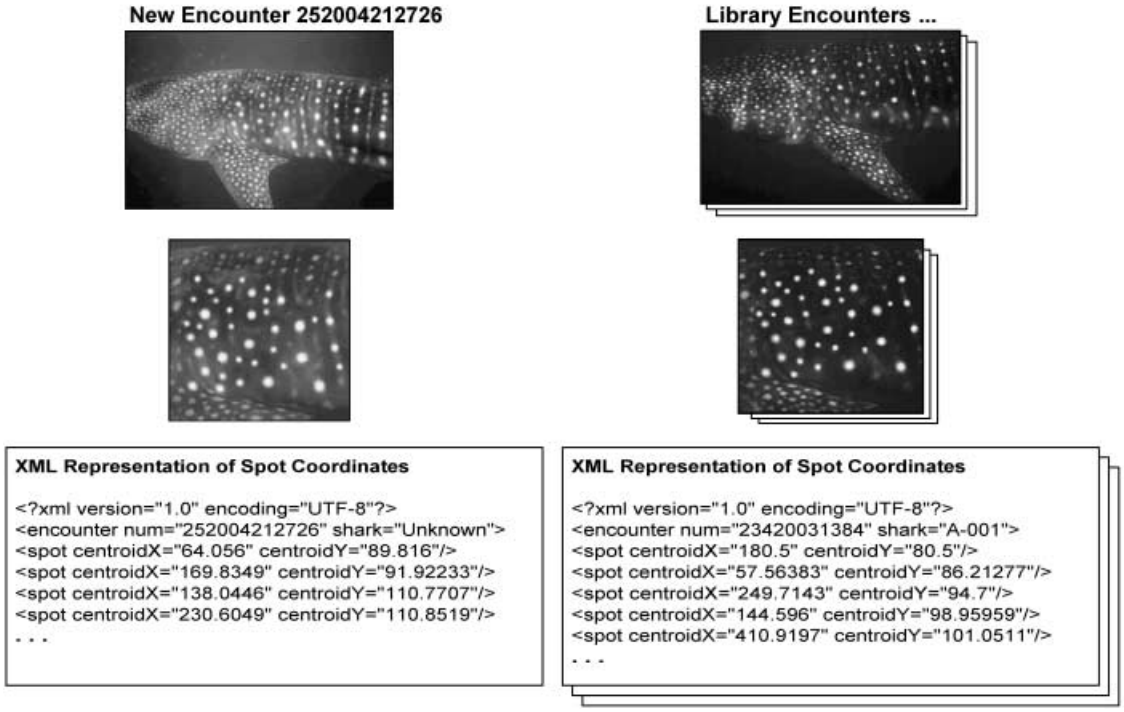


Fig. 1. Sample spot-pattern data sets from the ECOCEAN Whale Shark Photo-identification Library. Raw images (top row) from newly submitted (left) and catalogued (right) encounters are processed (see text) to highlight the naturally occurring spots (middle row), and a commercial software package is used to extract their coordinates within the image frame. The resulting lists of coordinates (bottom row) are then stored and input to the pattern-matching algorithm for identification and virtual 'tagging' of individual sharks.

and (iii) contact details of the submitter. The Library also served as the platform upon which our pattern-matching algorithm was developed and tested (Fig. 1).

While most of the raw data available for testing of our pattern-matching technique was collected by one of us (Norman 1999), submissions to the ECOCEAN Library from researchers, ecotourists, tour operators and others have been made, to date, from participants in 19 countries (see the Acknowledgements). Of particular importance has been the availability of sighting data spanning a 12-year period, 1992–2004, to confirm the reliability of spot patterning as a long-term identification tool.

To identify spot patterns, we select an area (the 'measurement region') located directly behind the gill slits on both the right and left sides of each shark. The region is bounded (i) anteriorly by the fifth gill slit, (ii) ventrally by the insertion plane of the pectoral fin, (iii) posteriorly by a line drawn vertically from the insertion point of the trailing edge of the pectoral fin and (iv) dorsally by the most ventral of the three longitudinal ridges (Fig. 2). This area can be easily photographed by a diver or snorkeller swimming alongside the shark. To ensure accurate photo-identification using the tools described here, photographers are encouraged to position their cameras as nearly as possible over the centre of the measurement region, with the field of view including both the vertebral column above and the pectoral fin below. Photographs of any secondary identification features are also encouraged, for example scarring on fins or body, that can be used to confirm the shark's identity.

SPOT EXTRACTION METHODOLOGY

A computer-driven pattern-matching system must be capable of clearly discerning features of interest in an image. The contrast of white whale shark spots on darker skin is well suited to a machine vision technique known as 'blob extraction', which measures the locations and dimensions of pixel groups of a single colour. The spatial relationships between these groups, represented by a set of derived (x, y) coordinates, form the basis for a unique identifier for each shark.

The variability of the underwater environment poses a challenge to feature recognition by introducing limiting factors such as low visibility and bright surface sunlight that may wash out spots. A whale shark may also be photographed with its anterior–posterior line forming an angle to the image horizontal. We compensate for these undesirable conditions through a series of image-processing steps.

Rotation correction

Using a graphics software package (Fireworks MX 2004, Macromedia, San Francisco, California, USA), the image is rotated until the segment of the shark's curved vertebral column directly above the measurement region is made parallel to a horizontal reference line (Fig. 2). The source image is then cropped along the boundaries of the measurement region.

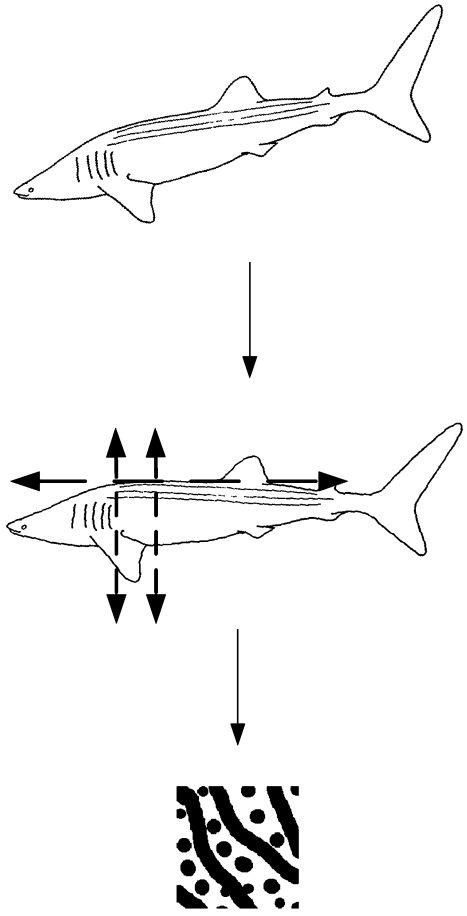


Fig. 2. Top: in a raw image submitted to the Library, the shark's orientation may be arbitrary. Middle: the image is rotated so that the vertebral column is made horizontal (long-dashed line) and the forward and rear boundaries of the measurement region (short-dashed lines) are vertical. Bottom: the image is cropped to isolate the correctly orientated pattern of spots and lines.

Contrast enhancement

Because blob extraction algorithms rely on a single colour to differentiate a blob from its local background, care must be taken to ensure that 'noise pixels' of that colour do not appear elsewhere in the image and cause false blobs, or in this case false white spots, to be counted and measured. This is especially true in images of whale sharks, which are most often photographed during daylight and near the surface, where reflected sunlight can produce white pixels in the source image. Artefacts of digital compression can also contribute spurious white pixels. False spots interfere with pattern matching and increase computation time by forcing additional calculations.

To reduce white pixel noise, Fireworks is first used to paint pure white spots on top of the natural shark spots, covering each with a best-fit circle. The contrast and brightness of the underlying image, but not of the painted white circles, are then reduced, increasing the overall contrast between the artificially superposed white spots and any noisy white pixels (Figs 1 and 7). The likelihood of extracting spurious spots is thus essentially eliminated.

After the photograph has been reduced to a cropped, rotation-corrected and contrast-enhanced greyscale image, spots are identified through blob extraction: a custom application was written using the eVision Easy-Object software library (Euresys, Angleur, Belgium) to determine the centre of gravity of each spot and to transmit a list of (x, y) coordinates via hyper-text transfer protocol (http) to the ECOCEAN Library. The list is stored in the Library as a matchable digital identifier associated with an encounter number and a set of photographs. The entire extraction process requires approximately 10 min for an experienced operator.

AN ASTRONOMICAL PATTERN COMPARISON ALGORITHM

Astronomers are frequently confronted with the task of identifying (and precisely locating within a coordinate system) stars, galaxies and other celestial objects in images of the night sky. Newly acquired images may be magnified, rotated or inverted relative to catalogued images of the same region, but the positions of objects common to both images can be used to derive the geometric relationship between the coordinate axes that underlie each image. A typical approach might locate common objects by identifying their surrounding patterns of stars.

Groth (1986) developed a pattern-matching algorithm for the comparison of two lists of coordinates, i.e. the (x, y) positions of stars, that effectively identifies individual points from one list with their likely counterparts in the other. The algorithm achieves the desired insensitivity to image magnification, rotation and inversion by forming triangles from selected triplets of coordinate points (Fig. 3). Geometrically similar pairs of triangles, one from each list, are then identified and a 'voting' process provisionally flags points that appear in multiple triangle pairs as being common to both lists. The method has been implemented as part of several astronomical data-reduction software packages, for example for the Hubble Space Telescope (STSDAS, Space Telescope Science Institute, Baltimore, Maryland, USA), and is cited in the literature (Schmidt *et al*. 1998). It has been demonstrated to be reliable even when the two lists of coordinates have as few as 25% of their points in common.

Here, we summarize the original algorithm's basic functioning, following Groth's (1986) notation. In the next section, we describe changes that we have made to optimize the method for use in identifying whale sharks, where the positions of stars are replaced by the coordinates of prominent spots in photographs of shark flanks.

Groth's (1986) triangle-based algorithm comprises the following steps. Hereafter, **A** refers to data derived from a newly acquired image and **B** refers to catalogued data. We assume, for this description, that coordinate lists **A** and **B** contain the same number (n) of points, but the algorithm does not require lists of equal length.

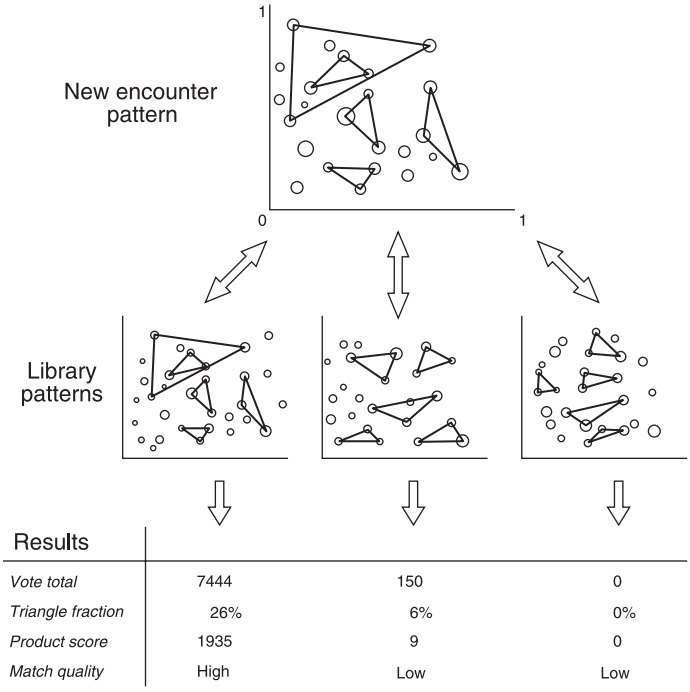


Fig. 3. A sketch of the basic pattern-comparison process based on the formation of triangles from triplets of points. Only subsets of all possible triangles are shown. Quantitative results are described in the text.

Filtering of coordinate lists

The coordinates of stars or spots (generically 'points') in each list are renormalized from their natural units, i.e. pixels, to the unitless interval [0, 1] while preserving the aspect ratios of the original images. A user-adjustable tolerance parameter (ε) is defined to quantify the typical uncertainty of coordinate measurements. To avoid confusion in pattern matching, the coordinates in each list are inspected to flag pairs of points that are too close together: separations less than a fixed multiple of the uncertainty (e.g. 3ε) are deemed too small and one of the points in the pair is purged from the list.

Formation of triangles

Every combination of three points within each coordinate list describes a triangle, with a point at each vertex. For **A** and **B** separately, all possible triangles are formed and their vertices indexed according to each triangle's shape: the shortest side is defined to lie between vertices 1 and 2, the intermediate side between vertices 2 and 3, and the longest side between vertices 1 and 3. The following geometric properties are then computed for each triangle, where (x_1, y_1), (x_2, y_2) and (x_3, y_3) are the coordinates of the indexed vertices.

The ratio of the longest (r_3) to the shortest (r_2) sides, $R = r_3/r_2$:

$$r_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad \text{eqn 1}$$

$$r_3 = \sqrt{(x_3 - x_1)^2 + (y_3 - y_1)^2}. \qquad \text{eqn 2}$$

The cosine of the angle at vertex 1:

$$C = \frac{1}{r_3 r_2}[(x_3 - x_1)(x_2 - x_1) + (y_3 - y_1)(y_2 - y_1)]. \qquad \text{eqn 3}$$

Tolerances in R (t_R) and C (t_C), assuming the coordinate measurement uncertainty ε to be independent in x and y and propagating this uncertainty through the expressions above:

$$t_R^2 = 2R^2F \qquad \text{eqn 4}$$

$$t_C^2 = 2S^2F + 3C^2F^2, \qquad \text{eqn 5}$$

where

$$F = \varepsilon^2\left(\frac{1}{r_3^2} - \frac{C}{r_3 r_2} + \frac{1}{r_2^2}\right)$$

is convenient shorthand and S is the sine of the angle at vertex 1.

The logarithm of the triangle's perimeter, $\log p$.

The orientation, i.e. whether the vertices 1, 2, and 3 are traversed in a clockwise or counter clockwise sense.

Filtering of triangles

For a coordinate list of length n, the number of triangles generated will be $n_t = n(n-1)(n-2)/6$. The results of the computations above are cumulated into new lists that record the properties of all n_t triangles for **A** and **B** separately. Figure 4 shows sample distributions of the R- and C-values for triangles derived from a whale shark spot pattern.

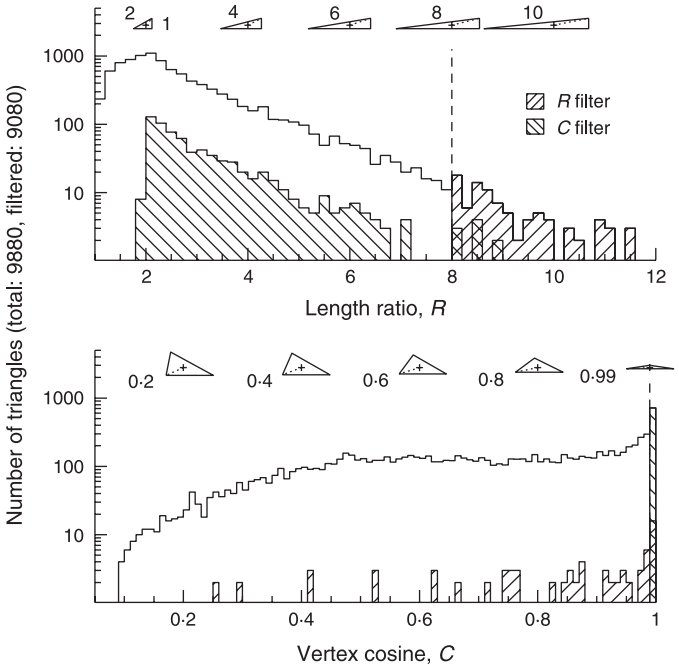


Fig. 4. Distributions of length ratios (R; upper panel) and cosines at vertex 1 (C; lower panel) for triangles derived from spot coordinates of the whale shark image in the left-hand panel of Fig. 1. The filtering criteria $R < 8$ and $C < 0{\cdot}99$ are indicated by dashed vertical lines in the upper and lower panels, respectively, and hatched regions show the resulting distributions of filtered triangles not suitable for matching. Along the top of each panel are triangles depicting representative geometries for R- and C-values on the abscissa. Crosses represent the triangle centroids, and dotted lines join each centroid to its vertex 1. In the upper panel, the triangles have vertical sides of unit length.

Not all triangles are well suited to pattern matching: some filtering is necessary. Triangles with large length ratios (we use $R > 8$ in Fig. 4) are discarded from both lists. Such elongated triangles produce large tolerances through equation 4; as a result, they can be falsely matched (see equations 6 and 7 below) with many dissimilar triangles, weakening the algorithm's ability to discriminate between different patterns.

Matching of triangles across lists

A given length ratio R and internal angle cosine C together describe a unique class of geometrically similar triangles within which triangles differ only in their relative size, i.e. by a magnification factor, and their orientations. At the heart of the pattern-matching algorithm, each **A** triangle's R- and C-values are compared with those for triangles from **B** according to matching criteria that depend on the tolerances t_R and t_C:

$$(R_A - R_B)^2 < t_{R_A}^2 + t_{R_B}^2 \qquad \text{eqn 6}$$

$$(C_A - C_B)^2 < t_{C_A}^2 + t_{C_B}^2, \qquad \text{eqn 7}$$

where both inequalities must be satisfied to declare a pair of triangles successfully matched. If more than one triangle from list **B** satisfies these criteria for a single **A** triangle, only the closest match, i.e. with the smallest value of the sum of the left-hand sides in equations 6 and 7, is retained.

For each pair of **A** and **B** triangles with similar geometry, the relative magnification factor (M) is computed:

$$\log M = \log p_A - \log p_B . \qquad \text{eqn 8}$$

If the **A** and **B** images contain the same point pattern, corresponding triplets of points will form many matching triangles all related by a common magnification factor. In contrast, any falsely matched triangles, i.e. **A** and **B** triangles that coincidentally have similar geometries but do not arise from the same triplet of points in the two images, will be related by an arbitrary magnification factor. True matches can therefore be distinguished from false matches by examining the frequency distribution of magnification values, e.g. the prominent peak at log M values near zero in Fig. 5 is dominated by true triangle matches, with a smaller contribution within the peak from the more broadly distributed false matches.

Similarly, the orientations (clockwise vs. counterclockwise) of member triangles among the matched pairs provide useful information. All true matches should have the same relative orientation, identical or opposite sense depending on whether the two data sets are mirror images of one another. In contrast, the set of false matches should reflect a random mix of same-sense and opposite-sense triangle pairs. This feature provides a rough estimate of the number of true, m_T, and false, m_F, matches found in the comparison set. If n_+ and n_- refer to the number of same-sense and opposite-sense matches, respectively, then:

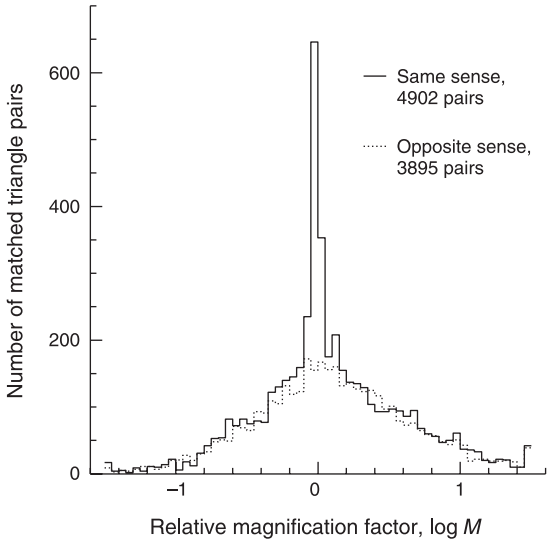


Fig. 5. Distributions of relative magnifications for pairs of geometrically similar triangles derived from the spot coordinate lists depicted in Fig. 1. Because the spot patterns are not mirror images, the excess of same-sense matches in the narrow central peak is evidence that portions of the two images contain the same point pattern. Opposite-sense (i.e. mirror-image) matches in this case are the result of chance occurrence and provide an estimate of the number of 'false' same-sense matches.

$$m_T = |n_+ - n_-| \qquad \text{eqn 9}$$

$$m_F = n_+ + n_- - m_T. \qquad \text{eqn 10}$$

To isolate the true matches, Groth (1986) describes a simple iterative filter that adapts itself to the log M distributions for matches of both senses. Our version of this filter is described in the next section.

Voting to identify points in common

At this stage, the algorithm has produced a number of matched triangle pairs, each of which involves three pairs of ostensibly matched vertex points. To determine which points are truly common to both data sets, it is assumed that matching points have a high probability of participating in more than one, probably many, matching triangles. This expectation is quantified through a voting scheme: every matched triangle pair casts three votes, one for each vertex pair. When all votes are cumulated, point pairs are ranked according to the number of votes they have received. If no pair receives more than one vote, the data sets are declared different. Otherwise, high-ranking pairs are assigned to one another as credibly matched points. These assignments continue until one of three conditions is met: the number of votes drops by a factor of two, a previously assigned point from either data set reappears in a different pair or, less commonly, the vote count drops to zero.

Iteration

Finally, the entire algorithm is run a second time, with input restricted only to those points that were matched in the first pass, to confirm or refute their associations.

THE SPOT-MATCHING ALGORITHM

We have tailored Groth's (1986) algorithm to reflect the properties of typical whale shark spot patterns and our data preparation and extraction procedures. These changes increase the algorithm's robustness but, at the same time, reduce its generality with respect to inversions and arbitrary rotations between the comparison data sets. We describe our changes, their motivations and their implications here.

Formation of triangles

We supplement the triangle properties considered by Groth (1986), R, C, their tolerances, log p, and orientation, with the following quantities.

A measure of each triangle's rotation relative to the image horizontal. The rotation angle is defined as a polar coordinate for vertex 1:

$$\theta = \tan^{-1}\left[\frac{y_1 - y_c}{x_1 - x_c}\right], \qquad \text{eqn 12}$$

where the origin (x_c, y_c) corresponds to the triangle centroid:

$$x_c = \frac{1}{3}(x_1 + x_2 + x_3) \qquad \text{eqn 13}$$

$$y_c = \frac{1}{3}(y_1 + y_2 + y_3). \qquad \text{eqn 14}$$

We adopt this 'local' measure of rotation over one that encompasses the whole image because it provides some insensitivity to distortions caused by the shark's curved body and projection effects in photographs acquired at oblique angles.

We quantify each triangle's size, s, adopting the fractional length of its longest side, r_3 in equation 2, relative to the maximum value $r_3^{\max}$ of any triangle in the image:

$$s = r_3/r_3^{\max}. \qquad \text{eqn 15}$$

Filtering of triangles

Along the flanks of whale sharks, and especially tailward of the pectoral fin, the distribution of spots typically becomes somewhat regular, falling along curved ventral–dorsal lines. Because the Groth (1986) algorithm forms triangles from all possible coordinate triplets, a number of flattened triangles (with C-values of nearly 1·0; Fig. 4) are generated in which all three vertices lie along a single arc. Such triangles from one image have a high probability of matching a large number of similarly 'flat' triangles from arcs in any arbitrary comparison

image. The anticipated sharp peak in the distribution of magnifications can then be diluted by the many falsely matched triangles. To suppress these unwanted false matches, we impose the constraint $C < 0{\cdot}99$ on triangles retained for analysis.

Projection effects related to the photographer's vantage point (see below) distort triangles, making them difficult to match. The distortion is greatest for triangles that span nearly the entire image, i.e. where the vertices lie near opposite edges of the measurement region. We therefore filter out these large triangles by requiring that $s < s_{\max}$, where tests show that a value $s_{\max} = 0{\cdot}85$ provides a good balance between rejecting distorted triangles and retaining useful ones.

Matching of triangles across lists

The triangle matching criteria in Groth's (1986) formulation, equations 6 and 7, are supplemented by a rotation criterion:

$$\theta_A - \theta_B < \theta_{\max}, \qquad \text{eqn 16}$$

where $\theta_{\max}$ is a user-selected parameter. The relative rotation between pairs of spot-pattern images is, by construction, small: we rotate each to align the shark's vertebral column with the horizontal axis. This information is used to match triangles; the rotational invariance of Groth's (1986) original algorithm, while useful for astronomical images, unnecessarily weakens pattern discrimination when both sets of spot coordinates are known to be based on the same coordinate system. If more than one pair of triangles is deemed a match according to equations 6, 7 and 16, the pair with the smallest quadrature difference δ is retained, where:

$$\delta^2 = \frac{(R_A - R_B)^2}{t_{R_A}^2 + t_{R_B}^2} + \frac{(C_A - C_B)^2}{t_{C_A}^2 + t_{C_B}^2} + \frac{(\theta_A - \theta_B)^2}{\theta_{\max}^2}. \qquad \text{eqn 17}$$

Similarly, the original algorithm's insensitivity to inversion of one of the images is not necessary: we assume that photographs submitted to the database are correctly orientated. We nevertheless track the number of opposite-sense triangle matches, n_-, in applying an iterative filter on the magnification factors. In each iteration, the mean and standard deviation of log M values are computed for same-sense triangles, and matches are discarded if they require magnifications more than z standard deviations from the mean value, where:

$$z = \begin{cases} 1, & \text{if } n_- > n_+ \\ 3, & \text{if } m_F < 0{\cdot}5 m_T \\ 2, & \text{otherwise.} \end{cases} \qquad \text{eqn 18}$$

Iterations continue until one of the following conditions is met: no matches are discarded in an iteration; no matches remain in the comparison set; the number of iterations reaches a pre-set limit of 20. If no matches remain, the two data sets are declared different and the algorithm terminates. Otherwise, all opposite-sense matches are discarded while same-sense matches are retained for voting.

Iteration and scoring of encounters

Voting for spot matches proceeds as in the original algorithm. A second pass through the entire code, with matched spots as input, effectively filters out any points incorrectly identified in the first pass. Our implementation departs from Groth's (1986) at this point in allowing single spots to be eliminated during the second pass without disqualifying the comparison pair of images as a potential match.

When two spot data sets are compared, a score is computed by summing the votes awarded to each pair of successfully matched spots: if v_i represents the number of votes cumulated for the ith pair of spots, the sum $V = \Sigma_{i=1}^{m} v_i$ terminates, in the typical case, when $v_{m+1} < v_m/2$. The vote total V is a useful measure of the similarity between the two input spot patterns. Comparisons across different data sets, i.e. whether the patterns in a pair of images are more closely matched than the patterns in a different pair of images, must, however, be interpreted with care, because the maximum possible score is not fixed; it is determined by the number of triangles in the smaller filtered data set. To account in part for this difference, the algorithm also reports the number of triangles that contributed votes for spots at their vertices (Fig. 3), as a fraction f_T of all available (filtered) triangles. We adopt as a final score for ranking purposes, S, the product of the vote total and this fraction of successfully matched triangles:

$$S = f_T V. \qquad \text{eqn 19}$$

We investigate, in the following section, the statistical properties of these three quantities to assess their utility in ranking spot-pattern comparisons.

We believe that this scoring scheme, together with the C and triangle-size filters that we have introduced, will prove useful in other applications of the Groth (1986) method. Table 1 summarizes the algorithm's adjustable parameters (e.g. coordinate tolerances and filtering criteria), the values recommended for astronomical images by Groth (1986), and the values we find provide robust performance for matching whale shark spot patterns. Optimized values were derived in most cases by examining the triangle properties of a handful of comparison pairs in detail, while others were derived by examining the scores of all visually confirmed matches as the parameters were varied. Users of the Library are provided the opportunity to alter these quantities to explore the algorithm's behaviour with different input images.

Results

Our spot-pattern matching technique was applied in database 'scans': as new whale shark photographs were

Table 1. Adjustable parameters of the triangle-based pattern-matching algorithm. Length units are normalized to the largest distance between two points in an image

Parameter	Adopted value		Description
	Groth (1986)	This work	
ε	0·001	0·01	One-dimensional coordinate uncertainty
R_{max}	10	8	Maximum triangle-side length ratio
C_{max}	NA	0·99	Maximum cosine of angle at vertex 1
s_{max}	NA	0·85	Maximum triangle size
θ_{max}	NA	10°	Maximum relative triangle rotation

NA, not applicable.

submitted to the ECOCEAN Library, spot data were extracted and compared with patterns from all previously submitted images, separately for left and right flanks. A list of candidate image matches was produced by the algorithm and ranked according to the computed score.

A subset of the Library entries, or 'encounters', represents multiple images of the same shark. As described below, these instances proved useful in estimating the method's self-consistency: if encounter A matches encounter B, and B matches encounter C, the technique should also provide a match when A is compared directly with C.

CORRECT MATCHES: THE METHOD'S EFFICACY

To explore the method's success rate and any potential difficulties, spot patterns for each of 21 previously identified (i.e. matched by eye) left-side images were scanned across all other available left-side spot data sets. As of 1 December 2004, there were 271 such data sets. Similarly, six known right-side images were compared within the catalogue of 181 right-side data sets. In the vast majority of cases, comparisons involving different sharks produced a zero score; in some cases, however, a small non-zero score resulted. We refer to the latter as false-positive matches. When the same shark was imaged in both encounters, a high score typically resulted, an outcome we refer to as a correct match. Occasionally, comparison of two same-shark images produced a low score, or a failed match. Figure 6 summarizes the results of these tests. The distributions of vote totals V, matched triangle fractions f_T, and product scores S resulting from comparison of the 27 previously identified pairs of encounters are shown in green. For the same set of comparisons, all false-positive match scores reported by the algorithm were accumulated; the resulting distribution is shown in red. A reliable method for identifying unique patterns should minimize the overlap in the red and green histograms. We find that for vote totals V (top panel of Fig. 6) the distribution of false positives is broad and encroaches, at the high end, on the vote totals garnered by the correct matches. An essential discriminator appears, however, in the triangle fraction (middle panel of Fig. 6): when f_T is restricted to values greater than 5%, the number of false-positive matches drops from 236 to 11, while just three correct matches are also flagged, two of which had, in any case, the lowest vote totals V. In the bottom panel (Fig. 6), the product score S incorporates the additional information contained in f_T. We find that the distribution of S for false-positive matches is well described by a log-normal that drops off rapidly for $S > 10$.

The available sample of previously matched pairs of encounters is small but, we believe, representative of the underlying statistical properties of correct and false-positive match scores. The results shown in Fig. 6 therefore suggest an empirical scheme for classifying the quality of a pattern match as scored by our algorithm.

A non-zero score S less than 10 is unlikely to represent a true match, but rather is characteristic of a false positive.

A score between 10 and 100, especially with a fraction f_T greater than 5%, represents a moderately strong likelihood that the two patterns under comparison are truly matched.

Any score above 100 represents a strong candidate for a correctly matched pair of spot-pattern images. The log-normal distribution of false-positive scores places the $S = 100$ boundary 3·6 standard deviations above the mean: this implies a formal probability of chance occurrence in this high-confidence category of better than 1 in 6000.

Based on these criteria, we can estimate a success rate for the method. From among the 27 previously identified pairs of encounters tested, 21 produced scores in the strong match category, another four in the moderately strong category, none were reported as weak candidates and two failed to match altogether. We combine the two higher-confidence categories to derive a success rate of 25 out of 27, or 92%. Although based on a small sample, this rate is encouraging and may well improve with time, as photographers mindful of the requirements of our technique strive to improve their vantage points in obtaining new photographs of whale sharks, as discussed below.

FAILED MATCHES AND FALSE POSITIVES: DIFFICULTIES ENCOUNTERED IN APPLYING THE METHOD

The performance of pattern-matching techniques is subject to factors beyond the control of any numerical

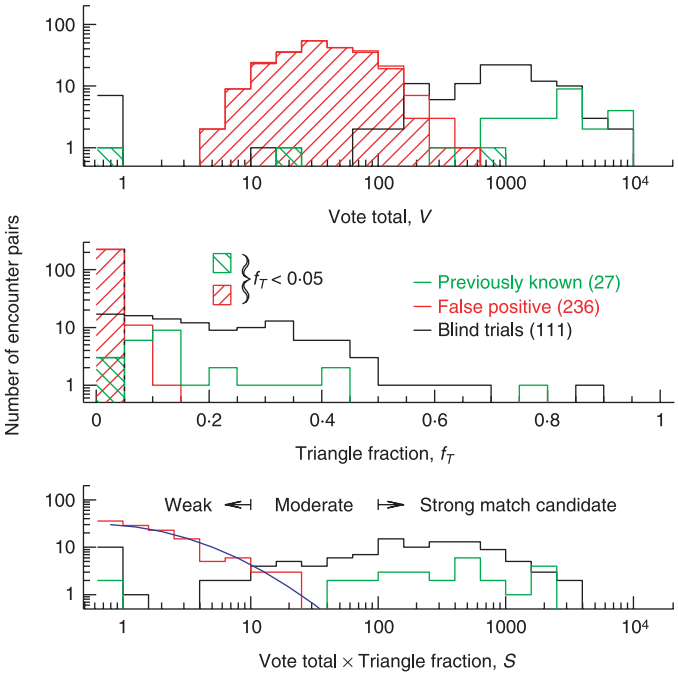


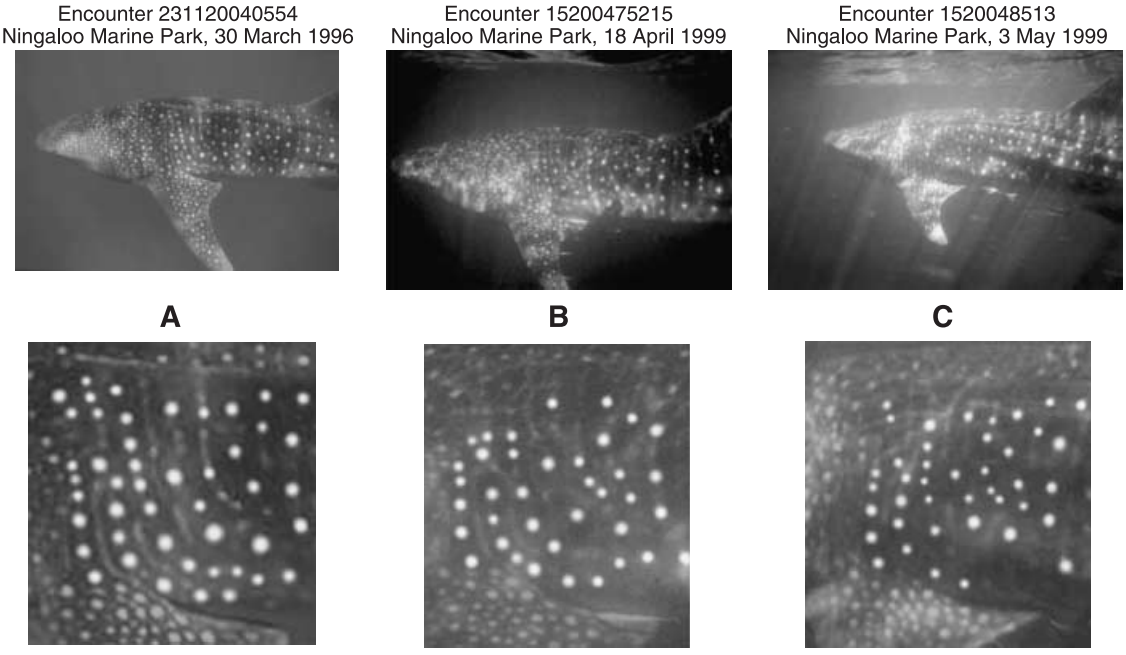
Fig. 6. Quantitative measures of match quality provided by the pattern-matching algorithm: vote total (top panel), fraction of triangles contributing votes (middle panel) and their product, our preferred ranking criterion (bottom panel). Right- and left-side trials have been combined. Distributions for correct (green) and false-positive (red) matches among previously identified images are shown, as well as those for new 'blind' matches (black) made by our algorithm. Trials resulting in zero votes are shown in the leftmost bin of the top and bottom panels. The mean (log $S = -0{\cdot}22$) and standard deviation ($\sigma_{\log S} = 0{\cdot}61$) of the false-positive scores are represented by a Gaussian curve in the bottom panel (blue). Hatched regions reflect trials in which fewer than 5% of triangles contributed to the vote total. A qualitative assessment of matches is suggested by the empirical scoring thresholds shown in the bottom panel, with weak, moderate and strong candidates corresponding to high, medium and low probability, respectively, of a false positive.

algorithm; our triangle-matching method is no exception. The difficulties that present themselves can be grouped into three categories: image quality, viewing geometry and spot pattern systematics.

Spot extraction from raw whale shark images can be complicated by lighting conditions, shadows, obscuration of spots by other fish, granularity of low-resolution images and other phenomena. Nevertheless, we find the triangle-matching algorithm to be effective even when two images have fewer than half of their spots in common, so that most of these difficulties are overcome simply by careful editing of photographs.

The direction from which shark flank images are obtained is important. In photographs obtained from directions anterior or posterior to the centre of the measurement region, foreshortening alters the aspect ratio of the spot pattern, changing the geometries of the derived triangles. Similarly, a camera vantage point too far dorsally or ventrally displaced produces altered geometries. The algorithm's ε uncertainty parameter can compensate, in part, for these distortions, and our implementation further mitigates perspective effects by imposing an upper limit on the size of triangles relative to the image dimensions. Nevertheless, an oblique image was responsible for one of the two instances in Fig. 6 in which a previously known match failed to produce a high score. We have experimented with numerical correction of spot patterns foreshortened along the anterior–posterior line by trigonometrically adjusting the spacings of spot x-coordinates immediately following extraction; although dependent on the operator's estimate of the angle formed between the image plane and the shark's flank, this technique holds some promise. We note that as the database of encounters grows, the collection of images for a given shark will span a range of perspectives, improving the odds that a successful identification will be made. As demonstrated in Fig. 7, photographs obtained from extreme forward or tailward angles will not be correctly matched with each other (simulations suggest that successful matches can be made for viewing perspectives different by up to 30°), but each will match other images made at intermediate angles. In the long term, therefore, oblique images of frequently encountered sharks will have minimal impact on the method's ability to provide a reliable identification.

As described earlier, whale shark spots sometimes fall along neatly arrayed arcs. Occasionally, spots are found to lie, within each arc, at quasi-regular intervals, so that they form a loose grid. When one image in a comparison pair exhibits such a gridded pattern, our algorithm can produce a relatively high score even when the images correspond to different sharks. Spots arrayed in grids account for the highest scores ($S \approx 20$) we have found among the false positives, and generally also produce $f_T > 0{\cdot}05$. Moreover, gridded patterns can



Comparison	A–B	A–C	B–C
Vote Total	3144	93	1589
Triangle Fraction	14·3%	0·3%	7·2%
Product Score	450	0	115
Match Quality	High	Low	High

Fig. 7. The effects of photographic perspective on scoring of numerical spot pattern comparisons. In the sequence of images A through C, the shark's head moves progressively away from the camera, so that image A is obtained from a vantage point essentially normal to the flank, while C's perspective is oblique. Contrast-enhanced spot patterns are shown in the lower row of images. Scores for the three comparisons quoted in the table demonstrate that adjacent pairs, with small angular displacements, produce reliable matches, but the comparison of A against C fails to produce a match because of distortion.

be responsible for failed matches, this is the case for the remaining failed match from our previously identified test data set, because falsely matched similar triangles from the two images overwhelm those that are correctly matched.

Where our algorithm fails to establish a strong match, visual inspection or some other method is needed to identify the imaged shark. False-positive outcomes are undesirable, but even in cases where the algorithm cannot provide an unambiguous identification, it reduces dramatically (by a factor of between 10 and 100) the number of images that a user need examine visually to uncover a successful match.

'BLIND' MATCHES: THE METHOD'S SUCCESSES

To date, 111 image pairs not previously known to be associated have been matched by our algorithm and, of these, 96 had scores $S > 10$. Typically, database scans produced a list of candidate matches, the most highly ranked of which were examined visually for spot-pattern compatibility and unrelated identification markers such as scars. Confirmed matches were noted and tabulated, resulting in the black histograms of score distributions shown in Fig. 6. As expected, most of the successful matches have scores in the high-confidence range, with decreasing numbers in the moderate- and low-confidence ranges. We note that not all of the blind matches constitute new identifications: in cases where three or more encounters were available for a single shark, all possible image pairs, for example three pairs for the shark shown in Fig. 7, were included in the category of blind matches, forming a rough self-consistency test of the method. The high-scoring fraction of 96/111 = 86% among blind matches provides supporting evidence for the method's efficacy. We emphasize that these results have been obtained with a data set that is not prejudiced against moderately oblique images; it reflects, in other words, a collection of encounter photographs that were acquired under real-world conditions.

Discussion

Computer-based recognition of natural spot patterning from digital (or digitized) images offers several new benefits. By solving the problem of scalability inherent in photo-identification by eye, a computer-aided method allows for 'data mining' of the large archive of images

acquired by researchers, management agencies, dive operators and tourists over the past two decades: our method has uncovered verifiable whale shark matches (from photographs as well as still frames captured from video footage) that pre-date the development of our algorithm. In essence, we have gone back and 'marked' a number of sharks that had not previously been physically or visually tagged, thereby increasing the number of sharks that can be 'recaptured' in the future. Already, the number of photographs and pattern samples in the ECOCEAN Library exceeds the ability of any single individual to match efficiently new photographs by eye. Rather, visual comparison now serves as a final validation of computer-executed scans that sift through hundreds of patterns with high accuracy in a short time. Built into the Web-based framework of the ECOCEAN Library, this system allows geographically dispersed researchers to make rapid identifications from a communal body of up-to-date data and research.

The implications of these capabilities for management and conservation may be profound. In several instances, for example for the southern right whale *Eubalaena australis* (Bannister, Kemper & Warneke 1996), conservation action plans already call for population studies through photo-identified mark–recapture analyses to determine whether threatened species form open or closed populations within well-defined geographical regions. For migratory animals such as the whale shark, the answer can determine whether conservation efforts should primarily be focused at the local or international level. Similarly, local researchers can use large photo-identification libraries over extended periods to determine the effectiveness of marine reserves in fulfilling their roles as protected sites for threatened species (Willis, Millar & Babcock 2003).

Identifying individuals repeatedly through photography can also inform biological observations, such as age of maturity, growth rate and foraging ecology. Among our algorithm's pattern-matching successes to date are high-scoring comparisons of images acquired 8 years apart (future submissions to the Library should allow matching across steadily longer time baselines). We find marginally significant evidence for a degradation in pattern-matching fidelity with increasing time spans, as might be expected if spot patterns evolve as sharks grow. It is possible that straightforward recognition of spot patterns will apply only to sharks larger than a certain minimum size, below which rapid growth in juvenile sharks may shift spot locations. To date, the algorithm has made successful multiyear matches with sharks as small as 4·5 m. A detailed study of this and other biological implications of new identifications will be presented elsewhere (B.M. Norman, J. Holmberg & Z. Arzoumanian, unpublished).

We have described a method for the automated identification of individual whale sharks from images of their spot patterning. These essentially unique and archivable digital 'fingerprints' can be used as natural markers to track individual fish over wide geographical areas and time spans much longer than can be achieved with other tracking techniques, provided that a large number of photographic encounters are organized and stored in a single repository. The ECOCEAN Whale Shark Photo-identification Library, created and maintained by the authors, serves this purpose. At the time of writing, the Library holds more than 1500 images, with more than 270 left-side and 180 right-side spot pattern data sets available for automated identification.

Although its performance is susceptible to degrading factors such as image quality, photographic perspective and the organized nature of spot patterns found on a small number of individuals, tests of the method using real-world data show that it identifies pairs of matched images with reliability nearing 90%, while producing a small number of false-positive matches that are easily discounted by visual inspection. The algorithm is thus a useful element in a toolbox of research technologies, such as satellite and data logging tags; for long-term population monitoring, virtual tagging eclipses plastic visual-identification tags, as these typically have a life span of less than 1 year.

We continue to work on refinements to the method and to explore the limits of its capabilities. Our implementation currently requires that a trained operator extract spot coordinates from submitted images and inspect the results of the automated scan across the image library. The latter is a desirable check on the method's scoring of image comparisons, but the former task can be further automated to improve efficiency and minimize the possibility of operator error. For example, a more sophisticated filtering scheme for triangle matches could restore the original algorithm's insensitivity to rotations of the image. We are also investigating techniques for extracting pattern information from the locations and shapes of lines that often accompany the spots on whale shark surfaces.

Acknowledgements

We are indebted to Dr G. Nelemans for bringing the Groth algorithm to our attention. We extend special thanks to the Western Australian Department of Conservation and Land Management, Australian Department of Environment and Heritage, ecotourism operators at Ningaloo Marine Park in Western Australia and in many other locations around the world, Murdoch University, the Thyne Reid Education Trust, the Australian Marine Conservation Society, the Fielman Foundation, Rhiannon Bennett, Suzy Quasnichka, Allison Richards, Ed Stastny and members of the general public who have assisted via submissions to the ECOCEAN Whale Shark Photo-Identification Library.

References

Baillie, J.E.M., Hilton-Taylor, C. & Stuart, S. (2004) *2004 IUCN Red List of Threatened Species: A Global Species Assessment*. IUCN, Gland, Switzerland and Cambridge. http://www.redlist.org/.

Bannister, J.L., Kemper, C.M. & Warneke, R.M. (1996) *The Action Plan for Australian Cetaceans*. Australian Department of the Environment and Heritage, Canberra, Australia. http://www.deh.gov.au/coasts/publications/cetaceans-action-plan/whaleap5a11.html.

Compagno, L.J.V. (1988) *Sharks of the Order Carcharhiniformes*. Princeton University Press, Princeton, NJ.

Evans, P.G.H. (2003) *EUROPHLUKES Database Specifications Handbook.* http://www.europhlukes.net.

Groth, E.J. (1986) A pattern-matching algorithm for two-dimensional coordinate lists. *Astronomical Journal*, **91**, 1244–1248.

Hammond, P.S., Mizroch, S.A. & Donovan, G.P. (1990) *Individual Recognition of Cetaceans: Use of Photo-Identification and Other Techniques to Estimate Population Parameters*. Special Issue No. 12. International Whaling Commission, Cambridge, UK.

Hillman, G., Wursig, B., Gailey, G., Kehtarnavaz, N. *et al.* (2003) Computer-assisted photo-identification of individual marine vertebrates: a multi-species system. *Journal of Aquatic Mammals*, **29**, 117–123.

Lapolla, F. (2005) *The Dolphin Project.* http://thedolphinproject.org.

Last, P.R. & Stevens, J.D. (1994) *Sharks and Rays of Australia*. CSIRO, Hobart, Australia.

Lettink, M. & Armstrong, D.P. (2003) An introduction to using mark–recapture analysis for monitoring threatened species. *New Zealand Department of Conservation Technical Series*, **28A**, 5–32.

Melville, R.V. (1984) Opinion 1278: The generic name *Rhincodon* A. Smith, 1829 (Pisces): conserved. *Bulletin of Zoological Nomenclature*, **41**, 215–217.

Norman, B.M. (1999) *Aspects of the biology and ecotourism industry of the whale shark* Rhincodon typus *in northwestern Australia*. MPhil Thesis. Murdoch University, Murdoch, WA, Australia.

Norman, B.M. (2000) *2000 IUCN Red List of Threatened Species.* IUCN, Gland, Switzerland and Cambridge.

Norman, B.M. (2004) *Review of the Current Conservation Concerns for the Whale Shark (*Rhincodon typus*): A Regional Perspective.* Coast and Clean Seas Project 2127 Final Report to the Australian Government. Department of the Environment and Heritage, Canberra, Australia.

Ormerod, S.J. (2003) Current issues with fish and fisheries: editor's overview and introduction. *Journal of Applied Ecology*, **40**, 204–213.

Schmidt, B.P., Suntzeff, N.B., Phillips, M.M., Schommer, R.A., Clocchiatti, A., Kirshner, R.P., Garnavich, P., Challis, P., Leibundgut, B., Spyromilio, J., Riess, A.G., Fillippenko, A.V., Hamuy, M., Smith, R.C., Hogan, C., Stubbs, C., Diercks, A., Reiss, D., Gilliland, R., Tonry, J., Maza, J., Dressler, A., Walsh, J. & Ciardullo, R. (1998) The high-Z supernova search: measuring cosmic deceleration and global curvature of the universe using type IA supernovae. *Astrophysical Journal*, **507**, 46–63.

Taylor, G. (1994) *Whale Sharks.* Angus & Robertson Publishers, Sydney, Australia.

Thompson, W.L., White, G.C. & Gowan, C. (1998) *Monitoring Vertebrate Populations*. Academic Press, Inc., San Diego, CA.

Urian, K. (2005) *Mid-Atlantic Bottlenose Dolphin Catalog.* http://moray.ml.duke.edu/faculty/read/mabdc.html.

Wilkin, D.J., Debure, K.R. & Roberts, Z.W. (1998) Query by sketch in DARWIN: digital analysis to recognize whale images on a network. Proc. SPIE Vol. 3656, *Storage and Retrieval for Image and Video Databases VII* (eds M.M. Yeung, B.-L. Yeo & C.A. Bouman), pp. 41–48. SPIE, Bellingham, WA, USA.

Willis, T.J., Millar, R.B. & Babcock, R.C. (2003) Protection of exploited fish in temperate regions: high density and biomass of snapper *Pagrus auratus* (Sparidae) in northern new Zealand marine reserves. *Journal of Applied Ecology*, **40**, 214–226.

Received 19 March 2005; final copy received 10 August 2005
Editor: Rob Freckleton