An Astronomical Pattern-Matching Algorithm for Automated Identification of Whale Sharks (*Rhincodon typus*)

Z. ARZOUMANIAN
Universities Space Research Association, NASA Goddard Space Flight Center, Greenbelt MD, 20770, USA
zaven@milyway.gsfc.nasa.gov
(Corresponding author. Tel: 01-301-286-2547/Fax: 01-301-286-1684)

J. HOLMBERG
3612 SE Stark St., Portland OR, 97214, USA
holmbergius@yahoo.com

B. NORMAN
ECOCEAN, c/o Centre for Fish and Fisheries Research, Murdoch University, South St., Murdoch WA 6150, Australia
ecocean@ozemail.com.au

Suggested running title: *Automated Identification of Whale Sharks*

Word count: 8369

Summary

1. The largest shark species alive today, whale sharks (*Rhincodon typus*) are rare and poorly studied. Directed fisheries, high value in international trade, a highly migratory nature, and generally low abundance make this species vulnerable to exploitation. Mark-and-recapture studies have provided our current understanding of whale shark demographics and life history, but conventional tagging has met with limited success. To aid in conservation and management efforts, and to further our knowledge of whale shark biology, an identification technology that maximizes the scientific value of individual sightings is needed.

2. We describe a novel technique for identifying individual whale sharks through numerical pattern-matching of their natural surface 'spot' colorations. Together with scarring and other visual markers, spot patterns captured in photographs of whale shark flanks have, in the past, been used to make identifications by eye. We have automated this process by adapting an 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 much 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 projection effects in photographs not ideally oriented with respect to the shark's flank.

4. We describe our implementation of the pattern-matching algorithm, estimates of its efficacy, its incorporation into the new Web-based ECOCEAN Whale Shark Photo-identification Library, and prospects for its further refinement. A subsequent paper (Norman et al., in preparation) will discuss in greater detail the biological and conservation implications of the capability to identify 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 for population and other studies through mark/recapture analyses. The pattern-matching software has been implemented through a Web-based system designed for the management of 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. Applications to other species are anticipated.

Key-words: conservation, population studies, marine and fisheries management, mark/recapture, whale shark

Introduction

The whale shark (*Rhincodon typus*) is one of approximately 370 species of shark alive today (Last & Stevens 1994), and is a member of the order Orectolobiformes, which are predominantly bottom-dwelling species, e.g. wobbegong and carpet sharks (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 species is regarded as rare, with as few as 320 sightings documented prior to the mid-1980s (Wolfson 1986). Many aspects of whale shark biology remain poorly studied, particularly with regard to life history and demographics. The World Conservation Union (IUCN) Red List of Threatened Species[1] lists the whale shark as vulnerable to extinction (Norman 2000).

A number of outstanding questions in whale shark ecology may be addressed through collection and subsequent collation of sighting data around the world (Norman 2004). Mark-and-recapture studies can be used in any situation where animals can be 'marked,' or otherwise identified, and 'recaptured,' or identified later by resighting (Lettink & Armstrong 2003). Analysis of resultant data can be used to estimate abundance, survival, recruitment, and population growth rates over time (Thompson et al. 1998). Importantly, this research can enable an improved assessment of the global conservation status of this species.

Whale sharks are born with unique body patterning on their skin that is retained throughout their lives (Norman 2004). Similar to a fingerprint in humans (Taylor 1994; Norman 1999), this natural patterning of lines and spots—specifically behind the gill slits—shows no evidence of significant change over years and can therefore be used to identify individual sharks. Through the combination of photographed encounters and spot-pattern matching, a shark may be 'tagged' without any 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.

While it may be possible to manage small numbers of shark identification photographs and distinguish individuals by eye, the process becomes inefficient and unreliable when collating data from many individual animals sighted in a large number of regions throughout the world. The availability of large quantities of data has rendered manual photo identification unfeasible, motivating instead the development of a Web-based library with the capability of scanning an entire database of encounter photographs in an automated way.

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,[2] an online database facility that archives digital images of whale sharks submitted by researchers and other interested parties. With a sophisticated pattern-matching capability and a growing library of images, a scientifically valuable number of individual sharks may be identified across wide geographic and temporal spans, enabling significant new advances in the study of whale shark life histories, migration

[1] http://www.redlist.org/

[2] http://photoid.whaleshark.org/

patterns, and demographics.

Materials and Methods

THE SHEPHERD PROJECT AND THE PROVENANCE OF PHOTO-ID 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 catalog framework that allows for the management of mark/recapture data accumulated by a global research community and interested third parties, such as ecotourists or government management agencies. This framework, which combines an object-oriented database, image management and protection functionality, an extensible programming interface, parameter search and multi-format data export capabilities, was completed in 2004 and first used in the ECOCEAN Whale Shark Photo-identification Library. The Library, built upon a J2EE software platform (Sun Microsystems, Inc., California, USA), is a repository for whale shark spot-pattern data and the photographs from which they are derived. Basic information required to accompany whale shark identification photographs includes a) date and location of the sighting, b) sex and size of the animal, and c) contact details of the submitter. The Library also served as the platform upon which our automated pattern-matching algorithm was developed and tested (Fig. 1). Its Web-based nature allows for worldwide access to data and the results of pattern-matching database 'scans,' which seek out and rank similarities in whale shark markings in a manner similar to an Internet search engine. Finally, the Library provides a data export facility to support trending and population analyses using software such as Microsoft Excel or Program Mark, allowing submitted data to become instantly useful to researchers and management agencies.

While most of the raw data available for the development of our pattern-matching technique was collected through the research of one of us (see Norman 1999), data submissions to the ECOCEAN Library from ecotourists, researchers, tour operators, managers, etc., have been made, to date, from participants in 19 countries where whale sharks have been sighted (see Acknowledgments). Of particular importance has been the availability of sighting data spanning a 12-year period, 1992–2004, especially to confirm the efficacy of spot patterning as a reliable long-term identification tool.

For the identification of individual sharks, we select an area (the 'measurement region') located directly behind the gill slits on both the right and left sides of the shark. The region is bounded as follows: a) anteriorly by the fifth gill slit, b) ventrally by the insertion plane of the pectoral fin, c) posteriorly by a line drawn vertically from the insertion point of the trailing edge of the pectoral fin, and d) dorsally by the most ventral of the three longitudinal ridges (Fig. 2). This is an area that can be easily photographed by a diver or snorkeller while swimming alongside the shark. To ensure accurate photo-identification using the tools described in this paper, underwater photographers are encouraged to position their cameras as nearly as possible over the center 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, e.g. scarring on fins or body that can be used to further confirm the identity of an individual shark. In addition to digital photographs and video, hard copy images can be scanned and subsequently submitted to the Library for analysis.

SPOT EXTRACTION METHODOLOGY

A reliable computer-driven pattern matching system must be capable of clearly discerning features of interest in the measurement region of an image. The contrast of white whale shark spots on colored skin is well suited to a common image manipulation process called 'blob extraction,' where, in this case, the blobs are the white spots to be distinguished from the darker background. The spatial relationships between the spots, as represented by a set of derived (x,y) coordinates, form the basis for the unique spot data pattern for each shark.

The variability of the underwater environment can compound the problem of computer-based feature recognition by introducing limiting factors such as low visibility or bright surface sunlight that may wash out spot patterns. Moreover, a whale shark may be photographed in configurations such that its anterior-posterior line forms an angle to the image horizontal. A variety of image processing techniques are available that can compensate for these undesirable environmental conditions, such as image rotation and sharpening of contrast by balancing color levels.

Rotation Correction. To standardize the shark's orientation before spot extraction, a straight, thin horizontal reference line is drawn, using a graphics package (*Fireworks MX 2004*, Macromedia, San Francisco, USA), over each image. The underlying image is then rotated until the segment of the shark's curved vertebral column directly above the measurement region is made parallel to the reference line. Finally, the source image is cropped down to the boundaries of the measurement region.

Contrast Enhancement for Noise Reduction. Because blob extraction algorithms rely on a single color and, in some cases, allow for varying shades thereof, to differentiate a blob from its local background, care must be taken to ensure that 'noise pixels' of that color 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 problematic in images of whale sharks because they are most often photographed during daylight and near the surface. Surface sunlight washing over the shark's body can cause white pixels to appear in the source image that may be mistaken for spots by the blob extraction software. Artifacts of digital compression can also cause spurious white spots. False spots can interfere with pattern matching and increase the computation time for processing 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 underlying contrast and brightness of the source image, but not of the painted white circles, are then reduced, thereby increasing the overall contrast between the artificially superimposed white spots and any noisy white pixels (see Figs 1 and 7). The likelihood of extracting spurious spots is thus essentially eliminated.

Once the digital source image has been reduced to a cropped, rotation-corrected, and contrast-enhanced grayscale image, the spot pattern can be identified by blob extraction software and stored in a database. In this case, a custom application was written using a commercial library (*eVision EasyObject*, Euresys, Angleur, Belgium) to measure the coordinates of the center of gravity of the spots in the processed image and to transmit these via the Internet to the ECOCEAN Whale Shark Photo-identification Library for storage as a matchable digital identifier. The entire extraction process requires approximately 10 minutes for an experienced operator.

AN ASTRONOMICAL PATTERN COMPARISON ALGORITHM

Astronomers are frequently confronted with the task of identifying (and precisely locating within a standard coordinate system) stars, galaxies, and other celestial objects that appear in images of the night sky. A typical technique involves comparing newly acquired images, which may be arbitrarily magnified, rotated, or inverted, with cataloged images of the same region of the sky—the positions of objects common to both images are used to derive the geometric relationship between the coordinate axes that underlie each image. With the advent of digital imaging and large machine-readable datasets, automated methods were needed to carry out these tasks, and in particular the difficult first step of identifying common objects. A useful approach might, for example, identify an imaged region of the sky by its characteristic pattern of stars.

Groth (1986) developed a pattern-matching algorithm based on the comparison of two lists of coordinates, i.e. the (x,y) positions of stars in astronomical images, 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. A second iteration of the technique using the provisionally identified points as input confirms or refutes their identification. The method, which has been implemented as part of several astronomical data reduction packages, e.g. the Hubble Space Telescope's STSDAS,[3] and is cited in the literature (e.g. Schmidt et al. 1998), has been demonstrated to be reliable even when the two lists of coordinates have as few as 25% of their points in common.

We have adapted this pattern-matching algorithm to the problem of identifying whale sharks, replacing stellar positions with the (x,y) coordinates of prominent spots in photographs of shark flanks. Here, we summarize the algorithm in its original form, following much of Groth's notation and describing the algorithm's basic functioning. In the next section, we describe changes that we have made to optimize the method for use with whale shark spot data.

Groth's triangle-based algorithm comprises the following steps. Hereafter, **A** refers to data derived from, e.g. a newly acquired image, and **B** refers to cataloged data. We assume, for the purposes of 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.

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, denoted ε, is defined to quantify the typical uncertainty of coordinate measurements; Groth suggests the unitless value $\varepsilon = 0.001$. To avoid confusion in pattern-matching, the coordinates in each list are then inspected to flag pairs of points that are too close together: any separation less than a fixed multiple of the uncertainty (e.g. 3ε) is 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

[3] http://www.stsci.edu/resources/software_hardware/stsdas/

longest side between vertices 1 and 3. The following geometric properties of all triangles are then computed, where (x_1,y_1), (x_2,y_2), and (x_3,y_3) are assumed to be the coordinates of vertices 1, 2, and 3, respectively:

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

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

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

- The cosine of the angle at vertex 1,

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

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

$$t_R^2 = 2R^2F \quad (4)$$

$$t_C^2 = 2S^2F + 3C^2F^2, \quad (5)$$

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 counterclockwise 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 the distributions of the R and C values of triangles derived from the whale shark encounter image in the left-hand panel of Fig. 1.

Not all triangles are well-suited to pattern matching and some filtering is therefore necessary. Triangles with large length ratios (Groth recommends $R > 10$, we use $R > 8$ in Fig. 4) are discarded from both lists. Such elongated triangles produce large tolerances through eqn 4; as a result, they can be falsely matched (according to the criterion, eqn 6, below) with many triangles, weakening the algorithm's ability to discriminate between different patterns. Although not required in Groth's original formulation, Fig, 4 shows a cutoff in the value of C beyond which we remove triangles from further analysis; we describe in the next section the motivation for this additional filtering criterion.

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,

$$(R_A - R_B)^2 < t_{R_A}^2 + t_{R_B}^2 \quad (6)$$

$$(C_A - C_B)^2 < t_{C_A}^2 + t_{C_B}^2. \quad (7)$$

where both inequalities must be satisfied to declare a pair of triangles successfully matched. This computationally intensive search for similar triangles across the two lists can be optimized by searching only a subset of **B** triangles for each **A** triangle, as described by Groth. Where the lists differ in length, we define list **A** to be the one with the smaller number of filtered triangles. 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 eqn 6—is retained.

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

$$\log M = \log p_A - \log p_B . \tag{8}$$

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. By 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 distribution of magnification values. Figure 5 shows the distribution of $\log M$ for the sample comparison depicted in Fig. 1. The prominent peak at $\log M$ values near zero (the expected value when both images have been scaled to unit linear dimensions) is dominated by true triangle matches, with a smaller contribution within this peak from the more broadly distributed false matches.

Similarly, the orientations (clockwise vs. counter-clockwise) 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 datasets are mirror-images of one another. By 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

$$m_T = |n_+ - n_-| \tag{9}$$

$$m_F = n_+ + n_- - m_T . \tag{10}$$

To isolate the true matches, Groth describes a simple iterative filter that adapts itself to the distribution of magnifications and relative orientations for all matches. In each iteration, the mean and standard deviation of $\log M$ values are computed, and matches are discarded if they require magnifications more than z standard deviations from the mean value, where

$$z = \begin{cases} 1, & \text{if } m_F > m_T \\ 3, & \text{if } m_F < 0.1\, m_T \\ 2, & \text{otherwise.} \end{cases} \tag{11}$$

Iterations continue until one of the following conditions is met: no matches are discarded in an iteration, no matches remain in the comparison set, or the number of iterations reaches a preset limit, e.g. 10. If no matches remain, the two datasets are declared to be different and the algorithm terminates. Otherwise, it is assumed that, if same-sense matches outnumber opposite-sense matches, there is no coordinate inversion between the two datasets, and the remaining opposite-sense matches are discarded. Alternatively, if $n_- > n_+$, coordinate

inversion is assumed and the remaining same-sense matches are discarded.

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 datasets, it is assumed that matching points have a high probability of participating in more than one, likely 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 datasets are declared to be 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 dataset 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. Groth's formulation strictly requires that all input points be successfully matched in the second pass as well.

THE SPOT-MATCHING ALGORITHM

We have tailored Groth's original astronomical algorithm to the task of identifying whale sharks, with modifications that reflect the properties of typical whale shark spot patterns and the data preparation and extraction procedures we use. 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 datasets. Building on the description of the original algorithm above, we describe our changes, their motivations, and their implications here.

Formation of triangles. We supplement the triangle properties considered by Groth—R, C, their tolerances, $\log p$, and orientation—with the following additional quantities:

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

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

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

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

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

In other words, rotation is determined by the angle formed between a triangle's median line for vertex 1 (the line joining vertex 1 with the triangle centroid) and the image horizontal axis. We adopt this 'local' measure of rotation over one that encompasses the whole image, e.g. the angle formed by the longest side of each triangle relative to the x axis, because it provides a degree of useful insensitivity to distortions of the image caused by the shark's curved body and projection effects in photographs not taken at right angles to the shark's flank.

- We quantify each triangle's size, s, simply adopting the fractional length of its longest side, r_3 in equations 1–2, relative to the maximum value $r_3^{\max}$ of any triangle in the image:

$$s = r_3 / r_3^{\max}. \tag{15}$$

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 algorithm forms triangles from all possible coordinate triplets, a significant number of narrow, flattened triangles (with C values of nearly 1.0) 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. In this circumstance, matched triangles provide little useful information for identification of a unique pattern—the anticipated sharp peak in the distribution of magnifications can be diluted by the many falsely matched triangles. To prevent these uninformative triangle matches from overwhelming correctly matched triangle pairs, we impose a constraint on the C-values of triangles retained for analysis of $C < 0.99$. Our tests have shown that this filter strongly suppresses triangles that produce unwanted false matches. As Fig. 4 demonstrates, our C filter does not overlap significantly with Groth's original R filter in the triangles that are disqualified.

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 $s_{\max} = 0.85$ provides a good balance between rejecting distorted triangles and retaining useful ones.

Matching of triangles across lists. The triangle matching criteria in Groth's formulation, equations 6–7, are supplemented by a rotation criterion,

$$\theta_A - \theta_B < \theta_{\max}, \tag{16}$$

where $\theta_{\max}$ is a user-selected parameter. The relative rotation between pairs of images in our spot-pattern database is, by construction, small: as described earlier, we rotate each image to align the shark's vertebral column with the horizontal axis. This information is used to match triangles—the rotational invariance of Groth's original algorithm, while useful for astronomical images, unnecessarily diminishes the method's effectiveness when both lists 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

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

Similarly, we do not require the original algorithm's insensitivity to inversion of one of the images: we assume that photographs submitted to the database are correctly oriented. We nevertheless track the number of opposite-sense triangle matches, n_-, in applying a variant of Groth's iterative filter on the $\log M$ distribution. In our version, the mean and standard deviation are computed only for same-sense triangles, with the multiplier z determined as follows:

$$z = \begin{cases} 1, & \text{if } n_- > n_+ \\ 3, & \text{if } m_F < 0.5 m_T \\ 2, & \text{otherwise.} \end{cases} \tag{18}$$

Iterations continue until no matches remain, none were eliminated in the last iteration, or 20

iterations are made. Following this procedure, which effectively isolates true from false matches, the remaining opposite-sense matches are discarded.

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 procedure 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 datasets 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 $V = \sum_{i=1}^{m} v_i$ 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 datasets, 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 specifically by the number of triangles in the smaller filtered dataset. To account in part for this difference, the algorithm also reports the number of triangles that contributed votes for spots at their vertices (see, e.g. Fig. 3), as a fraction f_T of all available (filtered) triangles. We adopt as a final score S for ranking purposes the product of the vote total and this fraction of successfully matched triangles,

$$S = f_T V. \tag{19}$$

As we show in the following section, a score value $S > 100$ is a reliable indicator that the algorithm has identified two instances of the same spot pattern, while $S > 10$ provides evidence of a potential match worthy of further review.

Table 1 summarizes the algorithm's adjustable parameters (e.g. coordinate tolerances and filtering criteria), the values recommended for astronomical images by Groth, and the values we find provide the most robust performance for matching whale shark spot patterns. Optimized values were found 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 behavior with different input images.

Results

Our spot pattern matching technique was applied in database 'scans': as new whale shark photographs were submitted to the ECOCEAN Library, spot data were extracted and compared to pattern data from all previously submitted images, separately for left and right flank images. A list of candidate image matches was then produced, usually less than ten in number, ranked according to the score computed by the algorithm. A subset of the Library entries, or 'encounters,' represented multiple images of individual sharks. 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 to C.

CORRECT MATCHES: THE METHOD'S EFFICACY

To explore the method's success rate and any potential difficulties, spot patterns for each of 27 previously identified (i.e. matched by eye) left-side images were scanned across all other available left-side spot datasets. As of Dec. 1, 2004, there were 271 such datasets. Similarly, eight known right-side images were compared to the remaining catalog of 181 right-side datasets. 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 very 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 the comparison of all 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): 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, 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 which 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, especially, 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 divers and researchers mindful of the simple requirements of numerical pattern-matching 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 algorithm; our adaptation of the Groth 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 acquired by divers can be complicated by lighting conditions, shadows, obscuration of spots by other fish, granularity of low-resolution images, and other phenomena. Nevertheless, the triangle-matching algorithm has proven effective even when two images have fewer than half of their spots in common, so that most of these difficulties are simply overcome by careful editing of photographs.

The direction from which shark flank images are obtained is important. In photographs taken from an angle away from the perpendicular, projection effects alter the apparent aspect ratio of the spot pattern, changing the geometries of the resulting triangles. Similarly, a camera vantage point too far above or below the center of the measurement region produces altered geometries. The ε uncertainty parameter in the triangle-matching algorithm can compensate, in part, for these distortions, and our implementation further mitigates the effects of an oblique image by imposing an upper limit on the size of triangles relative to the image dimensions—large triangles that span nearly the entire image will be most distorted and hardest to match. 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 oblique images by trigonometrically adjusting the spacings of spot coordinates in the x-direction 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 likely 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 (we estimate that successful matches can be made for viewing perspectives different by as much as 30°), but each will match other images made at intermediate angles. In the long view, therefore, oblique images of frequently encountered sharks will have a minimal impact on the method's ability to provide a reliable identification.

As described earlier, whale shark spots sometimes fall along curved, neatly arrayed arcs. In rare cases, spots are found to lie, within each arc, at quasi-regular intervals, so that they form a loose grid. We find that, when one image in a comparison pair exhibits such a gridded spot pattern, our triangle-based matching algorithm can produce a relatively high score even when the images correspond to different sharks. Such behavior is not surprising given the large number of geometrically similar triangles that can be produced from spots arranged in a loose grid. Spots arrayed in grids account for the very highest scores ($S \approx 20$) we have found among the false positives, and generally also produce $f_T > 0.05$. In addition to high false-positive scores, gridded spots may also be responsible for failed matches. Indeed, this is the case for the remaining failed match from our previously-identified test dataset, because falsely matched similar triangles from the two images overwhelm those that are correctly matched.

In instances 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 for a given submitted image, it reduces dramatically (by a factor of between ten and 100) the number of

images that a user need examine visually to uncover a successful identification.

'BLIND' MATCHES: THE METHOD'S SUCCESSES

To date, a total of 111 image pairs not previously known to be associated have been matched through the use of 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 Figure 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, e.g. 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 dataset that is not prejudiced against moderately oblique images; it reflects, in other words, a collection of encounter photographs that were acquired under real-world diving conditions.

Discussion

The use of natural spot patterning from digital (or digitized) images and of computer-based pattern recognition offers several new benefits to whale shark mark/recapture studies. A large number of photographs of whale sharks has been acquired by researchers, management agencies, dive operators, and tourists over the past two decades. Photo-based pattern recognition allows for 'data mining' of these archives—our method has uncovered verifiable matches among images (photographs as well as still frames captured from video footage) that predate the development of our algorithm. In essence, we have been able to go back and 'mark' sharks that had not previously been physically tagged, thereby increasing the number of sharks that can be 'recaptured' in the future. In addition, computer-based pattern recognition solves the problem of scalability inherent in photo-identification by eye. Already, we have found that the number of photographs and pattern samples in the ECOCEAN Library exceed the ability of any single individual to efficiently match new photographs by visual comparison, with each new photograph increasing the amount of time required for positive identification. Rather, visual comparison now simply serves as a final validation of computer-executed scans that can sift through hundreds of patterns with high accuracy in less than one hour. Built into the Web-based framework of the ECOCEAN Library, this automated system allows geographically dispersed researchers to access the most current data, and to make rapid identifications from a communal body of data and research.

An added benefit of the photo-identification methodology is that the research community can readily train tourists and dive operators to gather, during recreational encounters with whale sharks and while maintaining a safe and non-threatening distance from the sharks, photographs suitable for use in mark/recapture studies. Underwater cameras and housings are within the budgets of most divers, and the large quantity of data that could in principle be gathered through cooperation with the dive tourism industry can now be efficiently managed and processed using automated pattern-recognition and image management software such as those found in the

ECOCEAN Library.

Among the highlights of our algorithm's pattern-matching successes to date are high-scoring comparisons of images acquired as much as eight years apart. Future additions 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 may apply only to sharks larger than a certain minimum size, below which rapid growth may shift spot locations in juvenile sharks. To date, the algorithm has made successful multi-year matches with sharks as small as 4.5 meters. A detailed study of this and other biological implications of new identifications will be presented in a subsequent paper (Norman et al., in preparation).

Conclusion

We have developed a method for the automated identification of individual whale sharks from digital images of their spot patterning. The technique, adapted from an algorithm designed for comparing star patterns in astronomical images, is based on the geometric properties of triangles formed by joining all possible combinations of three spot coordinates within a well-defined measurement region. This essentially unique and archivable digital 'fingerprint' can be used as a natural marker to track individual fish over wide geographic 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 data repository. The ECOCEAN Whale Shark Photo-identification Library, created and maintained by the authors, serves this purpose. At the time of this writing, the Library holds over 1500 images, with more than 270 left-side and 180 right-side spot pattern datasets available for automated identification.

Although its performance is susceptible to degrading factors such as image quality, photographic perspective, and the highly organized nature of spot patterns found on a small number of sharks, 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, used to study whale sharks. For long-term population monitoring, virtual tagging eclipses plastic visual-identification tags, as these typically have a life span of less than one year.

We continue to work on refinements to the method to improve its robustness and ease of use, and to explore the limits of its capabilities, along several lines. Our implementation currently requires that a trained operator extract spot coordinates from submitted images (the procedure takes about 10 minutes), and inspect the results of the automated scan across the image library. The latter is an unavoidable, and ultimately desirable, check on the method's scoring of image comparisons, but the former task can in principle 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, in order to broaden the automated pattern-matching capabilities of the ECOCEAN Whale Shark Photo-Identification Library.

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

Compagno, L.J.V. (1984) *FAO species catalogue. Vol. 4. Sharks of the world. An Annotated and illustrated catalogue of shark species known to date,* FAO Fish Synopsis 125, Vol. 4, Pt. 1, Hexanchiformes to Lamniformes, Rome.

Groth, E.J. (1986) A Pattern-Matching Algorithm for Two-Dimensional Coordinate Lists. *The Astronomical Journal,* **91**, 1244–1248.

Last, P.R. & Stevens, J.D. (1994) *Sharks and Rays of Australia.* CSIRO, Hobart, 513 pp.

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*, 33 pp.

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

Norman, B.M. (2000) In: *2000 IUCN Red List of Threatened Species.* IUCN, Gland, Switzerland and Cambridge, UK. xviii+61 pp. (Book & CD). Also on http://www.redlist.org/

Norman, B.M. (2004) Review of the current conservation concerns for the whale shark (*Rhincodon typus*): A regional perspective. *Coast & Clean Seas Project 2127* Final Report to the Australian Government Department of the Environment and Heritage, 74 pp.

Norman, B.M., Holmberg, J. & Arzoumanian, Z. (in preparation). Whale shark photo-identification: Unique skin patterning 'tags' individuals.

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. *The Astrophysical Journal,* **507**, 46–63.

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

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

Wolfson, F.H. (1986) Occurrences of the whale shark, *Rhincodon typus*, Smith. *Indo-Pacific Fish Biology: Proceedings of the Second International Conference on Indo-Pacific Fishes.* Pp. 208–226 (eds. Uyeno, T., Arai, R., Taniuchi, T. & Matsuura, K.) Ichthyological Society of Japan, Tokyo.

Tables

Table 1: Adjustable parameters that govern the performance 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	This work	
ε	0.001	0.01	One-dimensional coordinate uncertainty
R_{max}	10	8	Maximum triangle-side length ratio
C_{max}	N/A	0.99	Maximum cosine of angle at vertex 1
s_{max}	N/A	0.85	Maximum triangle size
θ_{max}	N/A	10°	Maximum relative triangle rotation

Figures

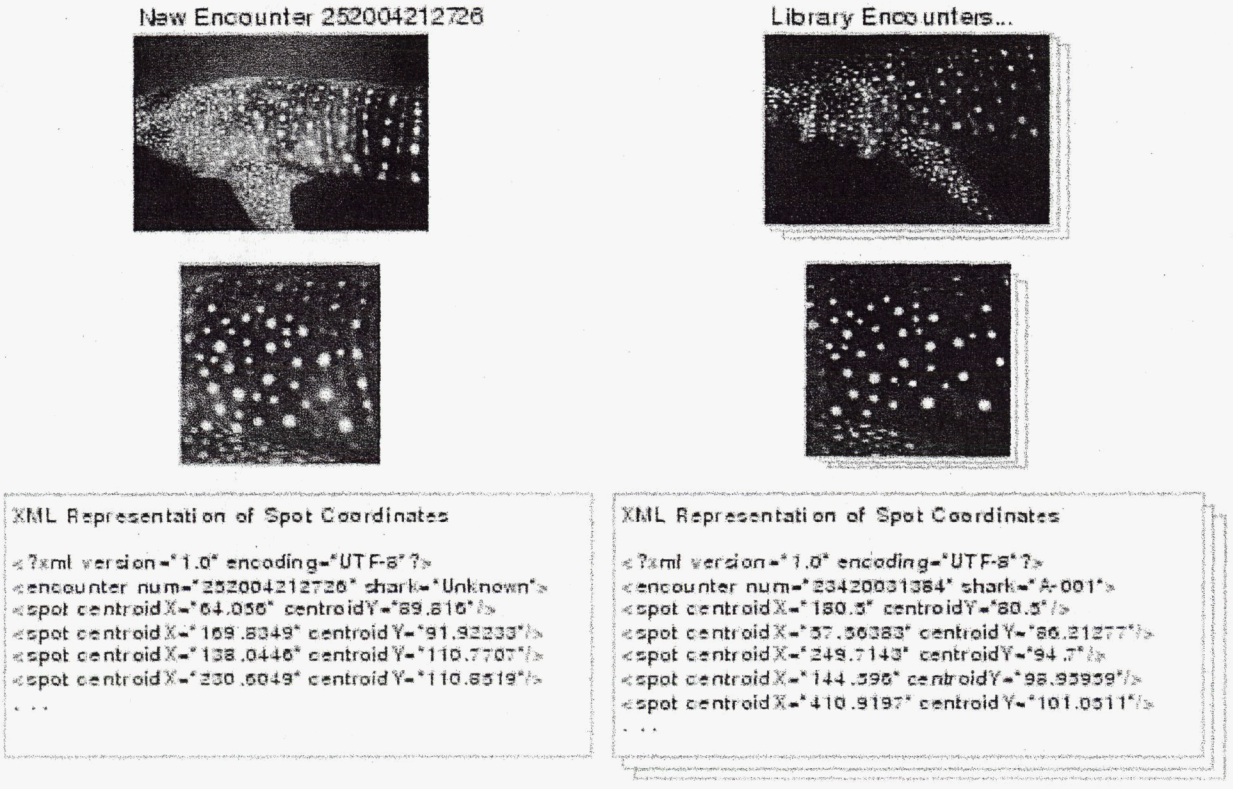


Figure 1: A sample of the ECOCEAN Whale Shark Photo-identification Library's spot pattern dataset. Raw images (*top row*) from newly submitted (*left*) and cataloged (*right*) encounters are processed (see text) to highlight the naturally occurring spots (*middle row*), and a commercial software package 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.

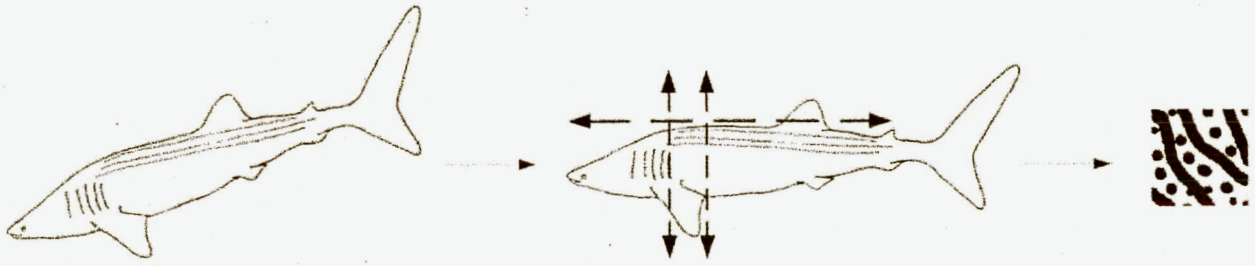


Figure 2: In a raw image submitted to the Library, the shark's orientation may be arbitrary. In these cases, the image is rotated so that the vertebral column is aligned with the image horizontal and the forward and rear boundaries of the measurement region are vertical. The image is then cropped to isolate the correctly oriented pattern of spots and lines.

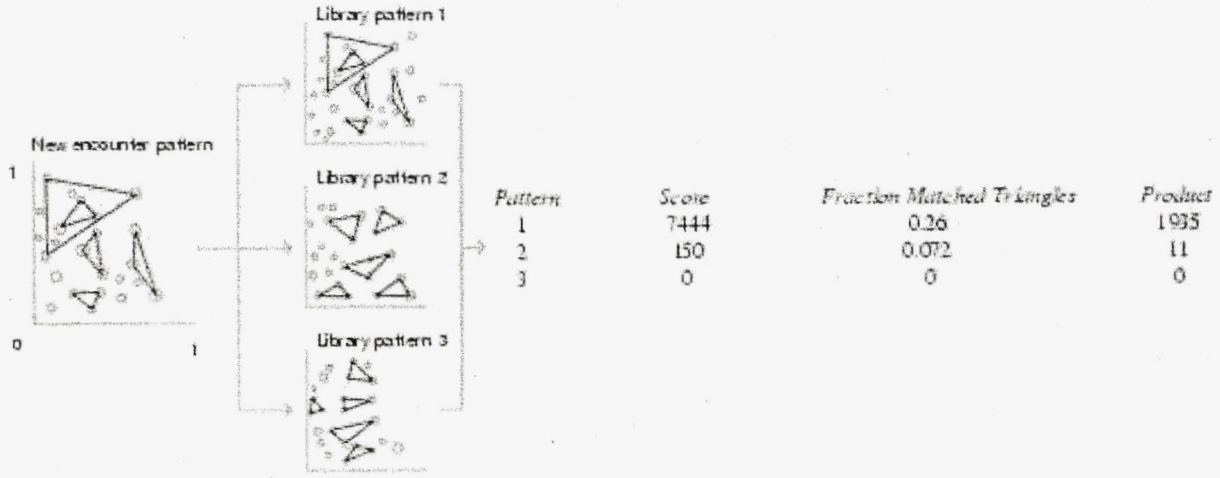


Figure 3: A sketch of the basic pattern-comparison process based on the formation of triangles from triplets of points.

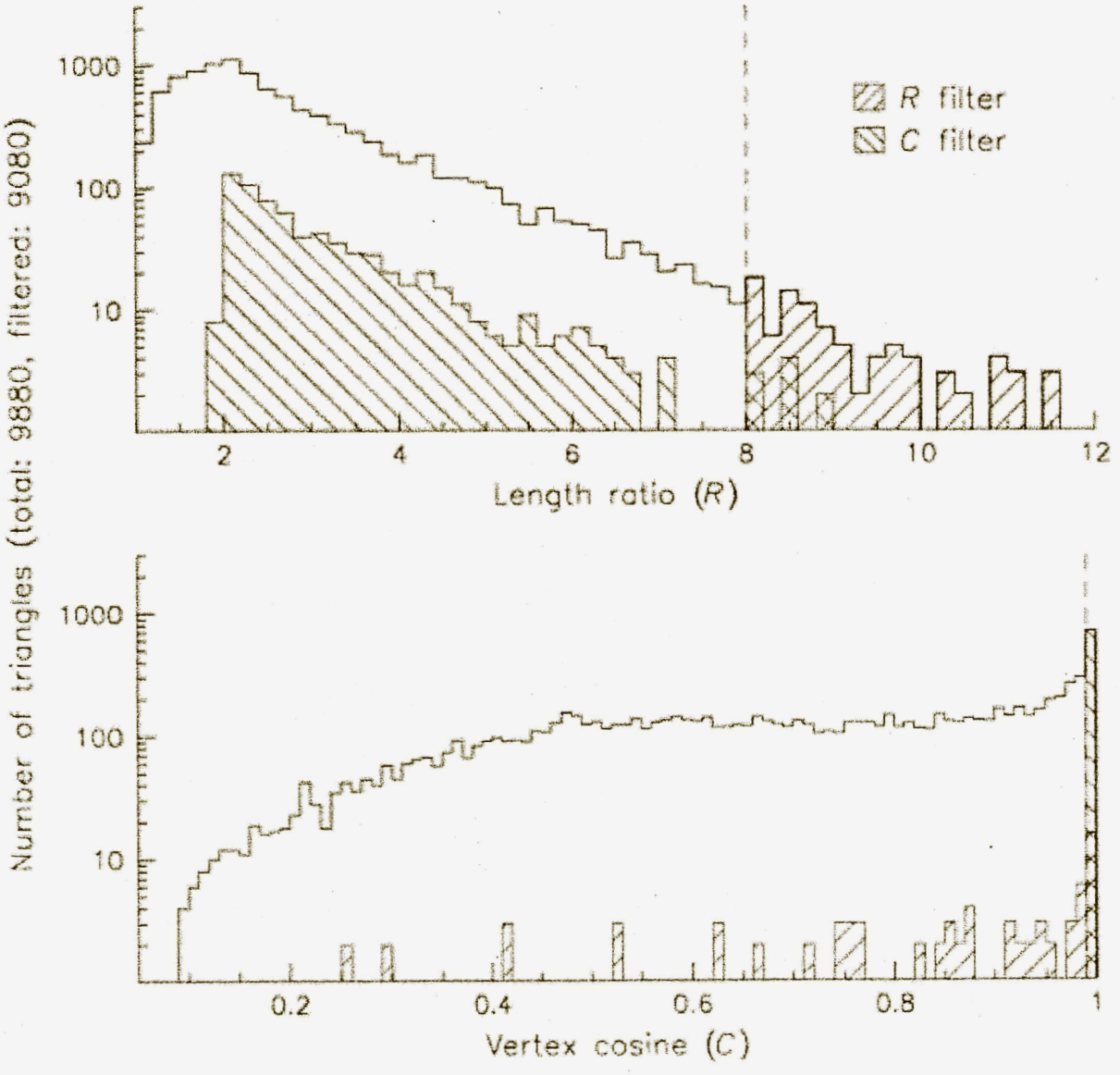


Figure 4: Distributions of ratio (R) and cosine at vertex 1 (C) values for triangles derived from spot coordinates of the whale shark image shown in Fig. 1. Note the logarithmic vertical axes. Filtering criteria for maximum R and C values are shown by dashed vertical lines, and hatched regions show the resulting distributions of filtered triangles not used for matching.

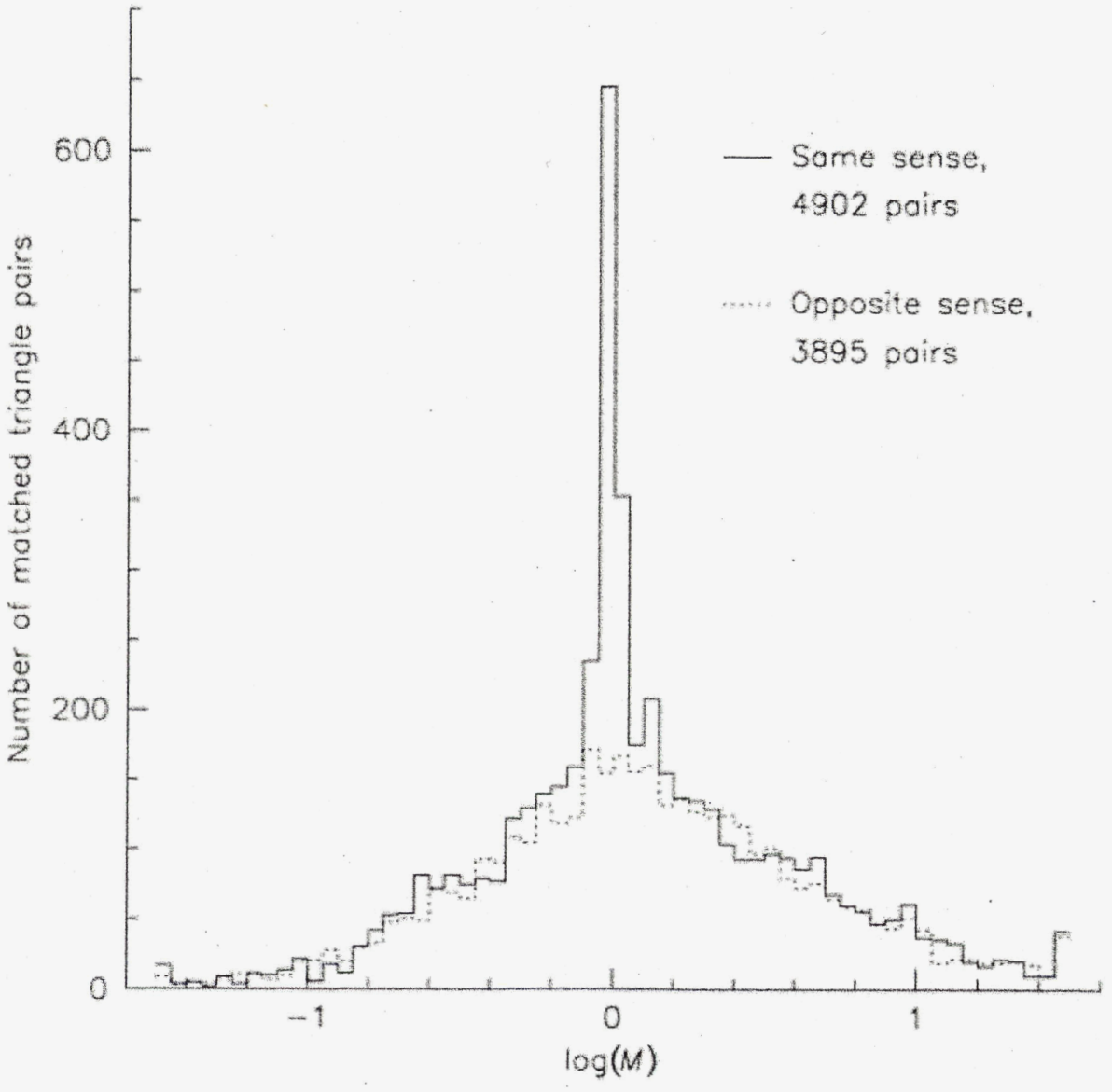


Figure 5: Distributions of magnifications for pairs of similar triangles derived from the spot coordinate lists depicted in Fig. 1. The narrow central peak for same-sense matches is evidence that portions of the two images contain the same point pattern. An iterative filter is used to isolate the matches contained in this peak.

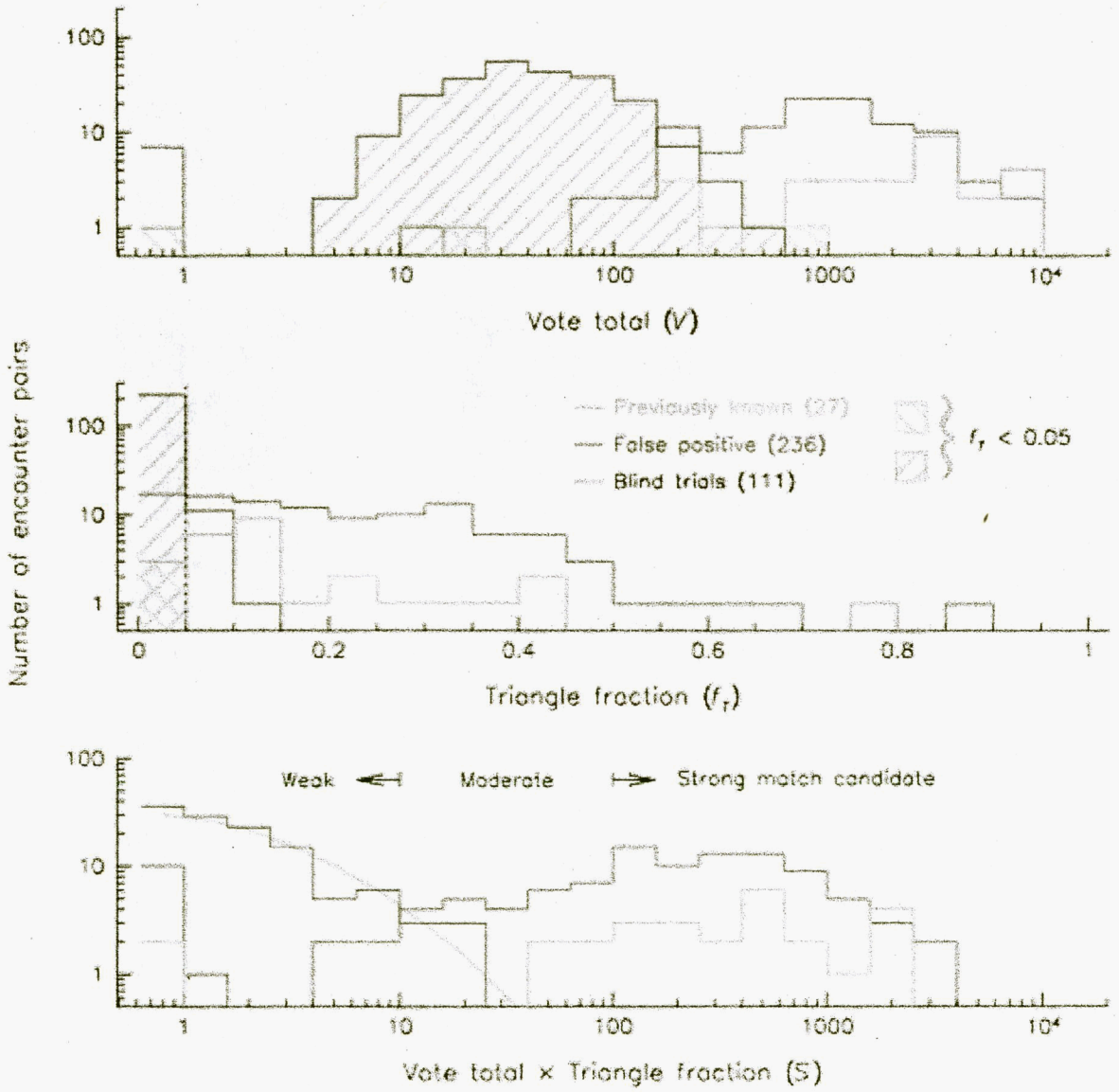
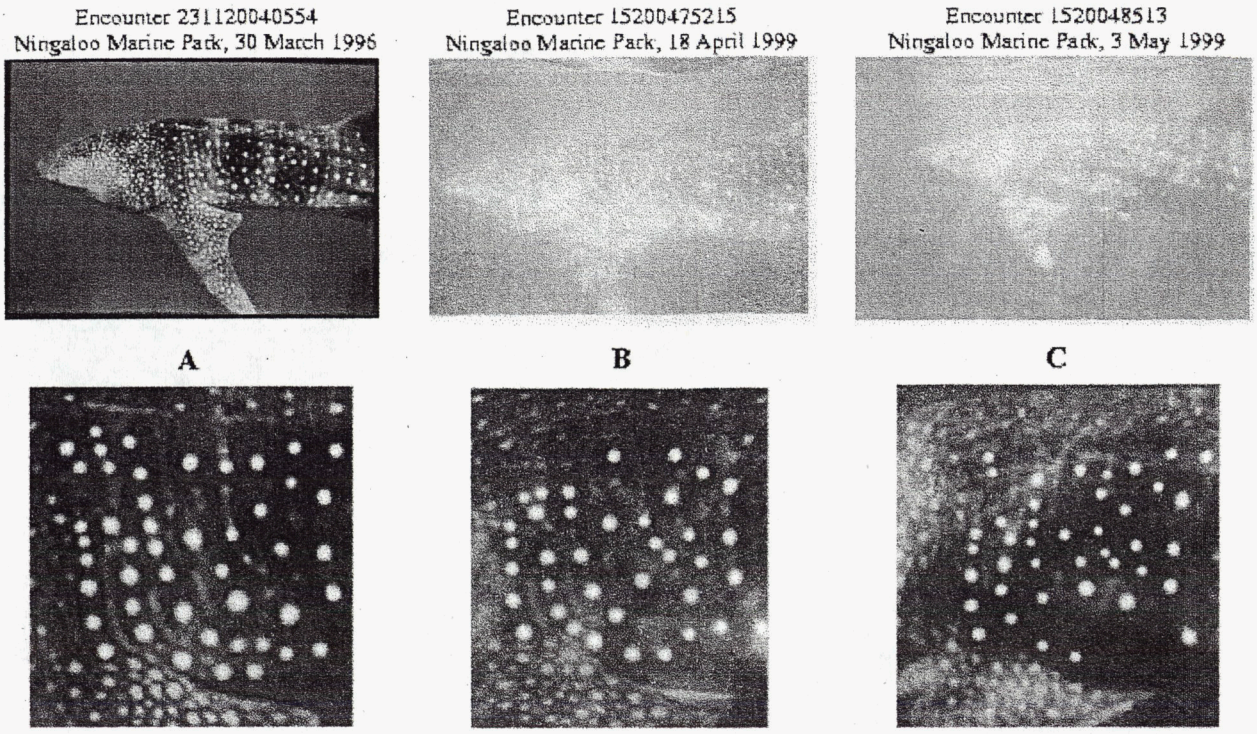


Figure 6: Quantitative measures of match quality provided by the pattern-matching algorithm: vote total (*top panel*), fraction of triangles contributing votes (*middle panel*), and their product, our preferred ranking criterion (*bottom panel*). Right- and left-side trials have been combined. Distributions 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 ($\log S = -0.22$) and standard deviation ($\sigma_{\log_S} = 0.61$) of the false-positive scores, in the log, 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.



Comparison	Vote Total	Triangle Fraction	Product Score	Match Quality
A – B	3144	14.3%	450.3	High
B – C	1589	7.2%	115.0	High
A – C	93	0.3%	0.3	Low

Figure 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.