Aquatic Mammals 2005, *31*(2), 243-247, DOI 10.1578/AM.31.2.2005.243

Comparison of Two Computer-Assisted Photo-Identification Methods Applied to Sperm Whales (*Physeter macrocephalus*)

Bas W. P. M. Beekmans,[1] Hal Whitehead,[2] Ruben Huele,[3]
Lisa Steiner,[4] and Adri G. Steenbeek[5]

[1,3]*Department of Industrial Ecology, Institute of Environmental Sciences (CML), P.O.Box 9518, 2300 RA, Leiden, the Netherlands*
[2]*Department of Biology, Dalhousie University, 1355 Oxford Street, Halifax, Nova Scotia, B3H 4J1, Canada*
[4]*Whale Watch Azores (WWA), 5 Old Parr Close, Banbury, OX16 5HY, United Kingdom*
[5]*National Research Institute for Mathematics and Computer Science, P.O. Box 94079, 1090 GB Amsterdam, the Netherlands*

Abstract

Two computer-assisted photo-identification methods for sperm whales (*Physeter macrocephalus*), namely the Highlight method (Whitehead, 1990) and the Europhlukes method (based on Huele et al., 2000), were compared. Performance was measured in terms of speed and accuracy. A test set was constructed containing two photographs of each of 296 individuals. The test set was divided into three classes of photographic quality and three classes of pattern distinctiveness. Both programs met requirements for rapid matching; the mean extraction times were 74.2 and 90.1 s per image for the Highlight and the Europhlukes methods, respectively. The two methods performed similarly with respect to accuracy. Accuracy improved by using higher-quality photographs or photographs representing more distinctive flukes. Still, even when using only the higher-quality photographs, 12.4% of the matches were not included in the top nine of the list of potential matches by the Highlight method compared to 14.0% for the Europhlukes method. The rate of failure to find the true match in the top nine was only 3.3% when both methods were used together, however. It is, therefore, recommended that for improved matching, both methods should be used in tandem or that an integrated program, which combines the two methods, should be developed.

Key Words: sperm whale, *Physeter macrocephalus*, photo-identification, computer-assisted matching, Highlight method, Europhlukes method

Introduction

Cetaceans are identified individually in three principal ways: (1) radio-tagging (e.g., Watkins et al., 1993, 1999, 2002), genetic-tagging (Palsbøll et al., 1997), and photo-identification (Defran et al., 1990). Whitehead & Gordon (1986) proposed photo-identification as a non-intrusive method of studying sperm whales (*Physeter macrocephalus*). Since then, photo-identification has been used to study many aspects of sperm whale biology such as horizontal movements (Whitehead, 2003), population sizes (Matthews et al., 2001; Whitehead et al., 1997), and social structure (Christal et al., 1998; Lettevall et al., 2002; Whitehead et al., 1991). Large sperm whale photographic catalogues have been constructed for several parts of the world. Because the difficulty of matching photographs increases with catalogue size, computer-assisted matching techniques are increasingly important.

Several automated photo-identification methods have been developed, including the Highlight method (as described by Whitehead, 1990) and the Europhlukes method (based on Huele et al., 2000). For both methods, a photograph is digitized, described, and compared against those already cataloged. Both methods use a matching algorithm, which computes a match coefficient (R-value) for each comparison (Whitehead, 1990). The output of both computer algorithms consists of an ordinal list of the best possible matches; the photographs with the highest R-values are at the top of the list. The researcher then checks the proposed matches visually, with the hope that, if there is a true match, it is near the top of the R-value list and if there is no match, most R-values are small.

The information concerning the trailing edge of the flukes used for matching differs between the two methods. For the Highlight method, the trailing edge is described in terms of the location of distinctive features such as nicks, scallops, and waves (Whitehead, 1990). The Europhlukes program (*Prototype*, *Version 1.2.1*) uses the whole contour for matching, as does FINSCAN (Araabi

et al., 2000), a third program not tested here. Although R-values range between zero and one for both programs, the scales are different. Earlier testing of the Highlight matching algorithm suggested that potential matches with an R-value ≥ 0.4 should be checked by eye (Whitehead, 1990). For the Europhlukes method, the suggested threshold R-value for visual checking is 0.6 (Steenbeek, pers. comm.). Therefore, R-values cannot be compared directly between the programs.

Herein, we present the results of a comparative test of the speed and accuracy of the Highlight and the Europhlukes methods.

Materials and Methods

Test Set

Negatives were used as source material. Almost all selected negatives were black & white and scanned at a resolution of 1,750 ppi by using a Nikon LS-2000 scanner. After scanning, the images were saved in tif-format. The test set (Table 1) consisted of 592 photographs, representing 296 matched pairs of 296 different sperm whales. The photographs originated from two sources, namely from Pacific sperm whales photographed by Hal Whitehead's laboratory at Dalhousie University (WL) and Atlantic sperm whales photographed by Whale Watch Azores (WWA). The known matched pairs from WL were all found originally by using the Highlight method. Using only these photographs for the construction of the test set could introduce a bias towards a more positive evaluation of the Highlight method. To correct for this possible bias, photographs from WWA also were included in the test set. The matched pairs belonging to the WWA collection were all found by the Europhlukes method. The test set consisted of 410 WL photographs and 182 WWA photographs.

The photographic quality of all photographs was determined by using the criteria of Arnbom (1987). Q-values were assigned to photographs in the WL catalogue during former projects (for instance, Dufault & Whitehead, 1995), so Q-values only needed to be assigned to the WWA photographs. The assigned Q-values were checked by an independent assessor. The photographs of each matched pair had the same Q-value, so Q-values could only differ between different matched pairs. Only photographs with quality Q ≥ 3 were included in the test set.

The distinctiveness of the trailing edge was represented by the number of marks, which was the average of the number of marks on the two photographs belonging to one matched pair, as assigned during application of the Highlight method. The pairs were divided into three distinctiveness classes: Class One: number of marks ≤ 10; Class Two: 10 < number of marks ≤ 20; or Class Three: number of marks > 20.

Table 1. Composition of the test sets used for matching photographs of sperm whales using the Highlight and Europhlukes methods; numbers represent photographs. Individual sperm whales were represented by two photographs of the same quality, as described by Arnbom (1987). WL= Whitehead Laboratory catalogue; WWA = Whale Watch Azores catalogue.

Source of photographs	Photographic quality			Total
	Q5	Q4	Q3	
WL	162	114	134	410
WWA	36	82	64	182
Combined	198	196	198	592

Analysis

The two matching methods were compared in terms of speed and accuracy.

Speed—The matching process could be divided into three different phases: (1) extraction and miscellaneous data input, (2) calculation of R-values by matching algorithm, and (3) visual check of suggested matches.

In this experiment, extraction and matching times were recorded for 150 photographs, which were processed by both methods. Time measurements for the extraction phase started when entering a previously digitized photograph and ended when the data were saved. Matching was carried out using the same computer (EMJ Academy PC with Intel Pentium III processor) for both methods. For the Highlight method, all photographs in the test set were matched against all others. The Europhlukes algorithm matched all photographs against all others, both the original and left-right reversed photographs (in case a photograph was reversed during digitizing, or taken of the dorsal side of the flukes). The corresponding total time of this matching was recorded for each method, and the mean time needed for processing a potential match was calculated.

Accuracy—Accuracy was measured by the proportion of false negatives, which was defined in this study as the probability of not finding the true match after checking the top n potential matches in the ordinal list of R-values for n = 1-9.

Results

Speed

Extraction Time—It took, on average, 72.4 s (SD = 15.76) to describe a trailing edge when using the Highlight method. The mean extraction time

for the Europhlukes method was 90.1 s (SD = 61.19).

Matching Time—It took 193.89 s for the Highlight matching program to match all 592 photographs with each other. The Highlight algorithm only matched photographs A with B, it did not match B with A, or A with itself. The number of matches executed by this program was (592 x 591/2) = 174,936 matches, resulting in an average matching time of 0.0011 s, (i.e., approximately 900 matches per s). Like the Highlight algorithm, the Europhlukes algorithm matched each photograph with each other. Furthermore, the Europhlukes algorithm matched each original photograph with all reversed photographs (except the reversed image of the original photograph). So, the Europhlukes algorithm matched photograph A with B, A with B-R, and B with A-R. It did not match B with A, A with itself, or A with A-R. Therefore, the Europhlukes program performed 592 x 591 = 349,872 additional matches, resulting in a total number of 524,808 matches. It took 398 s to perform this matching, so the average matching time per match equaled 0.00076 s (i.e., approximately 1,300 matches per s).

Accuracy

In Figure 1, the probability of a false negative is plotted against the number of potential matches checked. Results are shown for the whole test set and for two subsets, namely a subset consisting of Q=3 photographs and a subset consisting of Q=4 and Q=5 photographs. When the best 1 to 5 potential matches were checked, the Europhlukes method was a little more likely to find the true match than the Highlight method; however, when the best 6 to 9 potential matches were checked, the Highlight method performed slightly better. The probability of obtaining a false negative was considerably and consistently higher for both methods when using only Q=3 photographs. Even when only Q=4 and Q=5 photographs were used, 12.4% of the true matches were not included in the top nine of the list by the Highlight method, and 14.0% by the Europhlukes method. The Highlight method seemed to have more difficulties with twisted and tilted flukes: 36.1% and 19.5% of the true matches not included in the top nine of the list had these characteristics for the Highlight method compared to 19.5% and 12.2% for the Europhlukes method. The Europhlukes method seemed to have more problems with flukes which were largely smooth (41.7% for the Highlight method compared to 51.2% for the Europhlukes method) or had a part missing (2.8% vs. 26.8%).

Using both methods together resulted in consistently better performance (Figure 1). Only 3.3% of the matches were not included in the top nine of either list when matching Q=4 and Q=5 photographs. Photographs belonging to matches with higher fluke distinctiveness were matched more accurately with the two methods than photographs of less distinctive flukes. No large differences in accuracy were found between the two methods for each distinctiveness class, however. For instance, 48.4% of the matches belonging to Class One (n = 124 photographs) were not included in the top nine list generated by the Highlight method, compared to 42.7% for the Europhlukes method. These numbers were 17.5% and 21.5% for Class Two (n = 354) photographs and 7.0% and 5.3% for Class Three (n = 114) photographs. Furthermore, there seemed to be little bias in accuracy with respect to the origin of the photograph: matching performances of the Highlight and the Europhlukes methods did not significantly differ for both the WL and WWA photographs. For instance, 26.1% of the matches belonging to the WL catalogue were not included in the top nine list generated by the Highlight method, compared to 27.3% for the Europhlukes method. The top nine lists of the Highlight and Europhlukes methods did not include 12.6% of the matches within the WWA catalogue.

Discussion

Limitations of the Study

Two main problems exist for this comparison. Firstly, there was no ground-truth regarding the real number of matches in the test set. Known matched pairs were found by using either the Highlight or the Europhlukes methods. Therefore, it is possible that the test set consisted of more than 296 matched pairs. This seems to be a problem that cannot be circumvented when matches are based solely on photo-identification of one feature; double marking studies would be needed (e.g., Blackmer et al., 2000; Seipt et al., 1990). Even so, we believe that within the test set, all recorded matches were of the same whale, and that there were few if any unnoted matches present (see Dufault & Whitehead, 1995).

Secondly, the two methods were compared with respect to the matching of photographs. Matching of photographs is not only influenced by the method used, it also is influenced by photographic quality, fluke distinctiveness, and the user (Carlson et al., 1990). The whole test was executed by one user. We therefore neglected differences in user dependence during the testing, but the effects of photographic quality and fluke distinctiveness were considered. The results obtained by testing relied heavily upon the pictures selected. By selecting photographs with a range of quality and distinctiveness, we believe that we have circumscribed this problem.

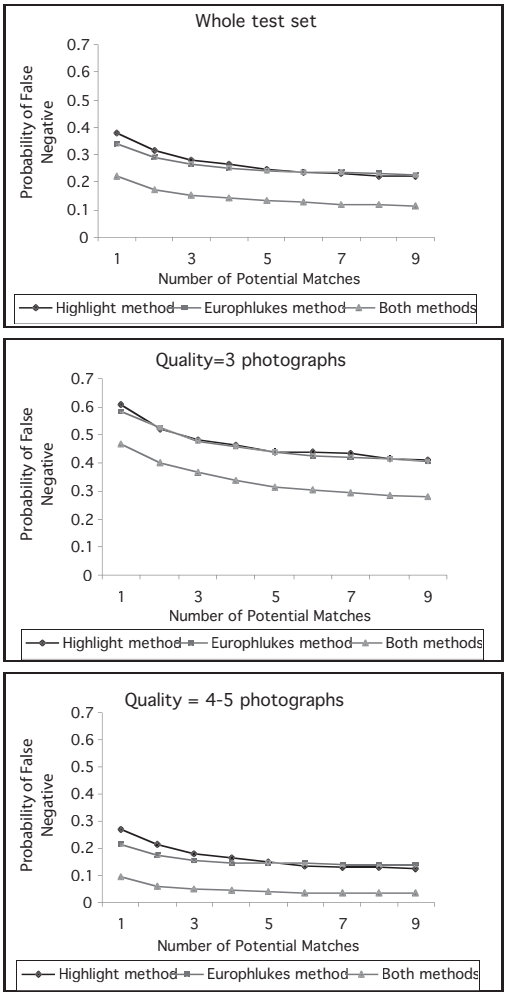

Figure 1. Probability of a false negative (i.e., true match not found); photo-identification of sperm whales by the number of potential matches.

Speed

In general, both the Highlight method and the Europhlukes method produced an average matching time that met the requirements for rapid matching, even with a large catalogue and a computer processor which was slower than those available today; however, it took longer for the Europhlukes method, compared to the Highlight method, to extract the information embedded in the trailing edge of the fluke. The principal reason that the Europhlukes program took longer was because of the size of the photograph. The Highlight method always shows the whole photograph on the monitor, whereas a large image is only partially displayed in the extraction program of Europhlukes. During extraction of large images in the Europhlukes method, considerable time was spent in dragging parts of the image onto the screen. This could be easily remedied in future versions of the program.

Accuracy

Photographic quality had a profound effect on the performance of the two methods. Photographic quality was negatively correlated with accuracy. This could be expected: the probability of missing or wrongly describing fluke marks is higher in lower-quality photographs. The inclusion of Q=3 photographs in the test set led to a decrease in accuracy. It is therefore advisable to either use only Q=4 and Q=5 photographs or to correct for, or at least consider, the false negatives that are likely to be introduced when Q=3 photographs are used for analysis. The higher the distinctiveness of the fluke, indicated by the number of marks in this study, the higher the accuracy in terms of ranking. This is to be expected, as flukes with more marks are less like other flukes.

An unexpected result of the test was the effect on accuracy of combining the two methods. Using both methods clearly improved the accuracy. In other words, the two matching methods complemented each other. For an improved accuracy, it is recommended to use both methods for photo-identification or to develop an integrated program of the two methods.

Other Differences Between the Methods

Photo-identification studies normally use databases consisting of photographs processed over a period of years, so it is quite probable that input for a matching program is generated by different users. The input for the Highlight program is more heavily influenced by the consistency between users. The user assigns descriptions to the marks and thereby decides whether the mark should be interpreted as a nick, scallop, or wave, for example. Sometimes, it is difficult to distinguish types of marks, like scallops or waves. For the Europhlukes program, the user defines the contour, without explicitly stating the nature of the marks. Therefore, more decisions are being made by the user when using the Highlight method, which increases the probability of differences in user input. Consistency in user input can be increased by a strong training program, which users need to do before having the authority to make final Highlight judgments.

Acknowledgements

We are especially grateful to Amanda Coakes for all of the advice and training she has given, thereby greatly contributing to a consistent use of the Highlight method. Furthermore, we would

like to thank all of the colleagues who have been involved in the taking and developing of photographs used in this experiment. Finally, we are grateful to two anonymous reviewers for their useful comments on the manuscript.

Literature Cited

Araabi, B. N., Kehtarnavaz, N., McKinney, T., Hillman, G., & Würsig, B. (2000). A string matching computer-assisted system for dolphin photoidentification. *Annals of Biomedical Engineering*, *28*, 1269-1279.

Arnbom, T. (1987). Individual identification of sperm whales. *Reports of the International Whaling Commission*, *37*, 201-204.

Blackmer, A. L., Anderson, S. K., & Weinrich, M. T. (2000). Temporal variability in features used to photo-identify humpback whales (*Megaptera novaeangliae*). *Marine Mammal Science*, *16*, 338-354.

Carlson, C. A., Mayo, C. A., & Whitehead, H. (1990). Changes in the ventral fluke pattern of the humpback whale (*Megaptera novaeangliae*), and its effects on matching: Evaluation of its significance to photo-identification research. *Reports of the International Whaling Commission*, *12*(Special Issue), 105-111.

Christal, J., Whitehead, H., & Lettevall, E. (1998). Sperm whale social units: Variation and change. *Canadian Journal of Zoology*, *76*, 1431-1440.

Defran, R. H., Shultz, G. M., & Weller, D. W. (1990). A technique for the photographic identification and cataloging of dorsal fins of the bottlenose dolphin (*Tursiops truncatus*). *Reports of the International Whaling Commission*, *12*(Special Issue), 53-55.

Dufault, S., & Whitehead, H. (1995). An assessment of changes with time in the marking patterns used for photoidentification of individual sperm whales, *Physeter macrocephalus*. *Marine Mammal Science*, *11*, 335-343.

Huele, R., Udo de Haes, H. A., Ciano, J. N., & Gordon, J. (2000). Finding similar trailing edges in large collections of photographs of sperm whales. *Journal of Cetacean Research and Management*, *2*(3), 173-176.

Lettevall, E., Richter, C., Jaquet, N., Slooten, E., Dawson, S., Whitehead, H., Christal, J., & Howard, P. M. (2002). Social structure and residency in aggregations of male sperm whales. *Canadian Journal of Zoology*, *80*, 1189-1196.

Matthews, J. N., Steiner, L., & Gordon, J. (2001). Mark-recapture analysis of sperm whale (*Physeter macrocephalus*) photo-id data from the Azores (1987-1995). *Journal of Cetacean Research and Management*, *3*(3), 219-226.

Palsbøll, P. J., Allen, J., Bérube, M., Clapham, P. J., Feddersen, T. P., Hammond, P. S., Hudson, R. R., Jørgensen, H., Katona, S., Larsen, A. H., Larsen, F., Lien, J., Mattila, D. K., Sigurjónsson, J., Sears, R., Smith, T., Sponer, R., Stevick, P., & Øien, N. (1997). Genetic tagging of humpback whales. *Nature*, *388*, 767-769.

Seipt, I. E., Clapham, P. J., Mayo, C. A., & Hawvermale, M. P. (1990). Population characteristics of individually identified fin whales, *Balaenoptera physalus*, in Massachusetts Bay. *Fishery Bulletin*, *88*, 271-278.

Watkins, W. A., Daher, M. A, Fristrup, K. M., & Howald, T. J. (1993). Sperm whales tagged with transponders and tracked underwater by sonar. *Marine Mammal Science*, *9*, 55-67.

Watkins, W. A., Daher, M. A, DiMarzio, N. A., Samuels, A., Wartzok, D., Fristrup, K. M., Gannon, D. P., Howey, P. W., & Maiefski, R. R. (1999). Sperm whale surface activity from tracking by radio and satellite tags. *Marine Mammal Science*, *15*, 1158-1180.

Watkins, W. A., Daher, M. A., DiMarzio, N. A., Samuels, A., Wartzok, D., Fristrup, K. M., Howey, P. W., & Maiefski, R. R. (2002). Sperm whale dives tracked by radio tag telemetry. *Marine Mammal Science*, *18*, 55-68.

Whitehead, H. (1990). Computer-assisted individual identification of sperm whale flukes. *Reports of the International Whaling Commission*, *12*(Special Issue), 71-77.

Whitehead, H. (2003). *Sperm whale societies: Social evolution in the ocean* (1st ed.). Chicago: The University of Chicago Press. 431 pp.

Whitehead, H., & Gordon, J. (1986). Methods of obtaining data for assessing and modelling sperm whale populations which do not depend on catches. *Reports of the International Whaling Commission*, *8*(Special Issue), 149-165.

Whitehead, H., Christal, J., & Dufault, S. (1997). Past and distant whaling and the rapid decline of sperm whales off the Galápagos Islands. *Conservation Biology*, *11*, 1387-1396.

Whitehead, H., Waters, S., & Lyrholm, T. (1991). Social organization of female sperm whales and their offspring: Constant companions and casual acquaintances. *Behavioral Ecology and Sociobiology*, *29*, 385-389.