Automated Intelligent Abundance Analysis of Scallop Survey Video Footage

Rob Fearn[1], Raymond Williams[1], Mike Cameron-Jones[1], Julian Harrington[2], and Jayson Semmens[2]

[1]School of Computing, University of Tasmania, Tasmania, Australia
[2]Marine Research Laboratories, Tasmanian Aquaculture and Fisheries Institute, Tasmania, Australia
{rcfearn, R.Williams, Michael.Cameron-Jones, Julian.Harrington, Jayson.Semmens}@utas.edu.au

Abstract. Underwater video is increasingly being pursued as a low impact alternative to traditional techniques (such as trawls and dredges) for determining abundance and size frequency of target species. Our research focuses on automatically annotating survey scallop video footage using artificial intelligence techniques. We use a multi-layered approach which implements an attention selection process followed by sub-image segmentation and classification. Initial attention selection is performed using the University of Southern California's (USCs) iLab Neuromorphic Visual Toolkit (iNVT). Once the iNVT has determined regions of potential interest we use image segmentation and feature extraction techniques to produce data suitable for analysis within the Weka machine learning workbench environment.

Keywords: Scallop Survey Video Transects, Automated Video Annotation.

1 Introduction

The Tasmanian Aquaculture and Fisheries Institute (TAFI) have been collecting underwater video of commercial and recreational scallop beds for over five years. The footage has been collected with the intention of developing minimally intrusive techniques for the quantitative assessment of scallop abundance and the impact of the habitat. TAFI currently use scallop dredging as their main methods of assessment of the commercial fishery but this technique can have a have a destructive impact on the marine environment and its inhabitants [1][2]. Dredge surveys also require the use of a commercial scallop vessel, whereas video surveys can be conducted on one of TAFI's own research vessels. Recently, the Tasmanian scallop industry have begun collecting their own underwater video in order to find new beds and determine the health of known beds, as they can do this all year round, including periods when by law they are not allowed to have their dredge on board the vessel. The drawback of the video approach is that footage can be collected (particularly by Industry) much

faster than it can be manually annotated and as a result TAFI's accumulating video footage has remained relatively untouched for a number of years.

Our research implements a multilayered approach to automatically annotating the scallop bed video footage captured by TAFI using Artificial Intelligence (AI) techniques. At this stage of our research the task being tackled is the identification and counting of commercial scallops from video footage of commercial scallop beds. We discuss the steps taken in video frame selection, analysis of conspicuous regions, segmentation, feature extraction and classification using the Weka Machine Learning Toolkit [3].

2 Video Footage Characteristics

The video footage used in this study varies in many aspects including environment, background colour and camera movement. TAFI's current underwater video system relies on a video camera being tethered to a vessel (drop camera) and is therefore affected by the speed the vessel is travelling at and the overall smoothness of the surface. The sea is seldom perfectly calm or flat, so the footage is subject to a constant up-and-down and rolling motion that is not perfectly uniform making it difficult to compensate for its effects.

TAFI have recently purchased a Remotely Operated Vehicle (ROV) and as a result it will be possible in the future to avoid the current undulation problem. However, for archived footage and footage collected by Industry vessels it is necessary to select sections of footage where the camera is deemed to be at an appropriate distance from the sea bed as footage from the AUV should be. When the camera is too high the scallops become ill-defined due to water clarity. When the camera is too low, individual scallops take over the entire frame. It is estimated that the preferred distance from the sea bed is approximately one third from the top of the camera's range of vertical movement in the current camera footage.

Fig. 1. A cropped sample of commercial scallop bed footage provided by TAFI.

Our primary focus in this study is on commercial scallop beds. Fortunately this footage involves comparatively sandy sea beds as opposed to some recreational scallop beds, for which some footage is also available, but there is less of a management need to examine this footage. Unfortunately, live scallops within the commercial scallop bed footage are usually partially buried under the sand, making it difficult to differentiate between sea bed and scallops. Thus annotation, whether manual or automatic, relies upon the somewhat darker shadow cast by the protruding perimeter of buried scallops which in many cases is crescent shaped (see Fig. 1).

3 Methodology

Our research breaks down the problem of counting commercial scallops into five main tasks. Frames from the video are first analysed by the iLab Neuromorphic Visual Toolkit (iNVT) to determine areas of potential interest which are then extracted as greyscale sub-images. Segmentation is performed on the sub-images to create a secondary binary sub-image. Feature extraction is then performed on both sub-images and this information is used to generate data in a suitable format for performing classification via the Weka Machine Learning Toolkit.

3.1 iLab Neuromorphic Visual Toolkit (iNVT)

To reduce the required search space within an image [4] we use the iNVT to identifying areas of potential interest prior to performing more complex segmentation and feature extraction techniques. iNVT is a software application developed by the University of Southern California (USC) and is designed to identify salient (conspicuous) regions within an image. Saliency is determined in a number of ways including intensity, orientation and colour.

iNVT generates a saliency map of the image and then uses a winner-takes-all neural network approach to determine the most salient region within an image [5]. It is possible to keep iNVT running for a specified number of attention shifts whereby salient regions are ignored once they have been *found.* iNVT outputs an ordered series of x and y coordinates specifying the centroid of each salient region found in the main image. Each set of coordinates is then fed into MATLAB and a 100 x 100 sub-image is extracted for further processing including segmentation and feature extraction.

3.2 Segmentation

Some of the feature extraction techniques used for classification allow us to use greyscale sub-images without the need to perform segmentation. We also create a dataset based on a binary version of each sub-image. Segmentation is a four stage process including blurring, contrast stretching, thresholding and clean-up.

The sub-image first has a Gaussian blur applied to it to help smooth the details of the sub-image and reduce the likelihood of visible interlacing lines caused from the extraction of still frames from compressed video. Blurring also helps to avoid regions with narrow channels or sections being split into two regions during thresholding.

The sub-image has its contrast stretched to help accentuate the typically darker region of shadow created by the edge of the scallop. Thresholding is then performed around the mean of the newly generated sub-image. The cleanup process is outlined in Fig. 2. We reduce the thresholded sub-image to either zero or one regions of potential interest based on the following rules:

1. All regions touching the sub-image boundary are removed as it is likely that a region located around the edge of the sub-image is not the area of interest that iNVT found.
2. Regions with too small an area (less than 60 pixels) are removed. Visual inspections of the binary sub-images has indicated that areas smaller than 60 pixels are generally featureless.
3. If more than one region still remains in the sub-image after the first two stages are performed the centroid for each region is calculated and the region with its centroid closest to the centre of the sub-image is chosen as the final region of interest. (The sub-image is generated around the point at which the iNVT determined a region of potential interest therefore the most central region should be most relevant).

It is important to note that it is possible that regions disposed of in the first and third steps of the cleanup process may still be picked up in another sub-image as it is possible for the iNVT to find two areas of interest within close proximity of each other. Further, when processing a sequence of frames, items not identified in one may be identified in others.

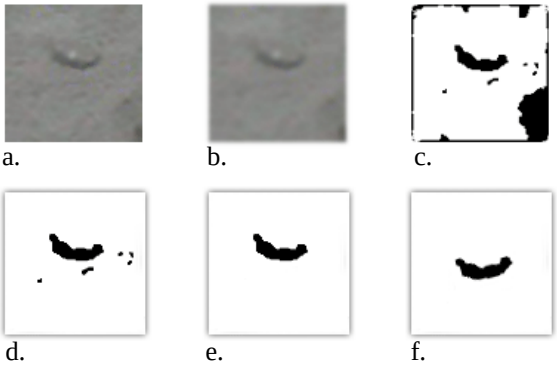
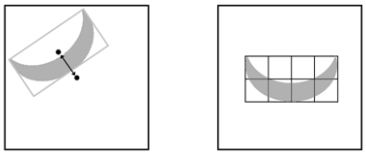


Fig. 2. a) Original sub-image. b) Gaussian blur and contrast stretch. c) Binary sub-image generated using thresholding around the mean. d) Regions touching the border of the sub-image are removed. e) Regions with an area < 60 pixels are removed. f) PCA may be applied to rotate and centre the remaining region around its major axis.

Erosion and dilation were also trialled on the sub-image's regions in an attempt to reduce the breaking up of regions prior to the cleanup process. However it was found that this process resulted in a loss of definition within the regions and provided little help rejoining regions that had been separated by the thresholding process. Ellipse fitting, whereby an ellipse is fitted to the lower margin of a region in an effort to define the that the scallop occupies, was also unsuccessfully tried as a potential final step in the segmentation process whereby an ellipse could be fitted to the underside of a region in an effort to find the potential space a scallop may occupy.

3.3 Feature Extraction

Various features have been extracted from both the greyscale and segmented binary sub-images. These include: seven invariant moments [6], the order in which an area of interest was initially selected by iNVT, the location and distance of the centroid of a region from the centre of the sub-image and the division of a region into segments (Fig. 3). When dividing a region into segments Principal Components Analysis is used to rotate the region so that its major axis is aligned with the x axis of the sub-image.



a. b.

Fig. 3. a) The Distance to Center is measured from the centre of the sub-image to the centroid of the region. X and Y coordinate information can also be extracted as a feature using the location of the region's centroid. b) The region is rotated and centered on its major axis using PCA. In this example the region has been divided using a 4x2 grid. The region's percentage of coverage in each division is then recorded to form a dataset.

Invariant moments were also extracted from the greyscale sub-image using a mask derived from the binary region. The mask was used at its original size and was also grown by varying amounts in order to encapsulate the surrounding areas of the region within the greyscale image. Results for all experiments are discussed in detail throughout section 4.

3.4 Instance Classification

We have consistently used 10 fold cross-validation with the same six classifiers from the Weka Machine Learning Toolkit to generate our results. We consider these classifiers will give us a reasonable representation of the overall performance of the system based on initial trials with other Weka classifiers. The classifiers are as follows:

- Multilayer Perceptron (MLP) - (a Neural Network)
- Naïve Bayes (NB)
- IB1 (Single Nearest Neighbour)
- Multiboost AB (MB-AB)
- NBTree (NBT – Naïve Bayes Tree)
- Decision Table (Dec T)

4 Experiments and Results

It is necessary to manually label each sub-image as either a scallop or non-scallop for system evaluation. This process is an arduous task and we have as yet not had the opportunity to ground truth the data. Subsequently, the dataset undoubtedly includes some misclassified instances; however, we are confident that enough instances have been classified correctly to give a clear indication of the system's overall performance. Table 1 and 2 outline how the dataset is affected by the segmentation and feature extraction processes. It also highlights the affects of various sub-image sizes ranging from 30x30 to 100x100. The *exceeds perimeter* row denotes the number of instances that exceed one or more of the original image's perimeters. These instances cannot be extracted as a complete sub-image and are therefore discarded.

100x100 was chosen as the most appropriate size for a sub-image without further tests being performed on larger sub-images. As shown in Table 1, by increasing the size of the sub-image beyond 100x100 we will at most avoid excluding a maximum of four extra scallop regions. At the same time we will reduce the number of useable instances by about 30 if we consider the average increase in the loss of useable instances across the *exceeds perimeter* row. We consider scallop instances that have been lost in the cleanup process after extraction of the sub-image to be misclassified as they will be excluded from any quantitative assessments we make of commercial scallops within the video footage.

Table 1. The effect the sub-image size and cleanup process have on the total number of scallop instances available for training and testing.

	30x30	40x40	50x50	60x60	70x70	80x80	90x90	100x100
Exceeds perimeter	76	119	136	172	205	232	262	295
Region on border	624	431	233	105	46	22	13	4
Area < 60	113	194	234	202	134	93	51	47
Training/Testing	14	83	224	348	442	480	501	481

Table 2. The effect the sub-image size and cleanup process have on the total number of non-scallop instances available for training and testing.

	30x30	40x40	50x50	60x60	70x70	80x80	90x90	100x100
Exceeds perimeter	293	414	525	619	730	815	900	1005
Region on border	622	428	227	121	61	33	18	8
Area < 60	1752	1691	1594	1384	1092	881	678	519
Training/Testing	32	166	353	575	816	970	1103	1167

The results presented below represent the findings we have made whilst testing various stages of our multilayered system. All results are presented as the percentage of correctly classified instances in Weka with the Standard Deviation given in parentheses. We have at all times worked with as many instances from the original dataset as possible. A total of 1648 instances (481 scallops and 1167 non-scallops) remain after the cleanup process is complete. Some of the cleanup process experiments resulted in greater losses of instances. In these cases the number of scallop and non-scallop instances classified by Weka has been written next to the input column description. When considering the fact that non-scallop instances account for approximately 70% of the total instances, this sets a minimum baseline level of performance that a classifier must exceed to be at all useful.

The results in Table 3 were generated to test the performance of the system using invariant moments as the only feature extraction technique. By feeding the invariant moments from both sub-images into Weka the average overall performance of the system is improved by approximately 5%.

Table 3. Invariant moment feature extraction using the greyscale and binary images and a combination of both. All sub-images are 100x100.

Sub-image	MLP	NB	IB1	MB-AB	NBT	Dec T	AVG
Grey & Binary	76 (3.7)	71 (3.1)	71 (5.1)	71 (1.7)	77 (3.8)	76 (3.3)	**74**
Binary	73 (1.9)	61 (3.3)	67 (3.8)	71 (1.0)	73 (1.7)	70 (2.2)	**69**
Grey	71 (2.4)	62 (3.1)	67 (2.3)	71 (0.1)	71 (3.7)	74 (3.4)	**69**

Table 4 outlines the impact of not discarding instances when the segmented region has an area less than 60 pixels. The addition of these instances increases the occurrence of non-scallops to approximately 75% and ultimately reduces the overall performance of the system compared to the best average result of 74% in table 4. This confirms the findings from our visual inspections that scallop instances with an area less than 60 are likely to lack features suitable for classification against non-scallop instances.

Table 4. Performance of grey and binary sub-image when no area rule is applied to the cleanup process resulting in a greater number of instances for classification.

Sub-image - No Area Rule	MLP	NB	IB1	MB-AB	NBT	Dec T	AVG
No Area, Grey & Binary (523/1247)	78 (2.3)	69 (4.2)	74 (3.2)	76 (1.8)	77 (1.6)	77 (1.8)	**75**

Table 5 demonstrates that by including a distance to centre (DTC) measurement from the centroid of the segmented region to the centre of the sub-image and the X and Y coordinate of this centroid we can marginally improve the performance of the system compared to the results in Table 4.

Table 5. Invariant moments from greyscale and binary sub-images with the inclusion of a distant to centre (DTC) measurement and the X and Y coordinate of the segmented regions centroid.

Sub-image	MLP	NB	IB1	MB-AB	NBT	Dec T	AVG
DTC	78 (2.7)	72 (3.0)	73 (5.0)	71 (0.4)	78 (4.1)	79 (3.9)	**75**
DTC, X&Y	78 (3.4)	74 (2.8)	73 (4.2)	71 (0.4)	77 (4.9)	79 (3.1)	**75**

Applying a mask to the greyscale image (using the segmented binary region) does little to improve the overall performance of the system. The mask is grown by a specific amount (outlined in Table 6) and is used as a means of encapsulating features in close proximity to the segmented region in the binary sub-image. The system performs at its best when the mask is grown by 10 to 15 pixels and a slight advantage can be gained on by including the distance to centre measurement and the coordinates of the region's centroid. However this advantage is marginal and when we compare the results of the individual classifier learning methods in Tables 5 and 6 we see that the best performing classifiers in Table 5 have their performance reduced in table 6 suggesting that although the average performance is similar the addition of a mask into our system offers no real advantage.

Table 6. Invariant moments using the segmented image as a mask over the greyscale image.

Sub-image mask	MLP	NB	IB1	MB-AB	NBT	Dec T	AVG
5x5	76 (2.3)	70 (3.5)	73 (4.9)	74 (4.7)	74 (5.1)	74 (4.2)	**74**
10x10	77 (2.9)	72 (4.2)	74 (3.4)	75 (3.2)	76 (3.1)	76 (2.8)	**75**
15x15	79 (1.7)	73 (2.9)	72 (3.5)	76 (2.9)	77 (3.3)	76 (3.0)	**76**
20x20	77 (2.7)	71 (2.3)	74 (3.5)	74 (3.0)	77 (2.6)	76 (1.8)	**75**
15x15, DTC, X & Y	78 (3.4)	73 (2.7)	75 (3.7)	75 (3.2)	78 (3.8)	78 (4.4)	**76**

The results presented in Table 7 outline the system's performance on the classification of the segmented binary sub-image when its region is divided in *n x m* sections. This approach outperforms all other results presented in this paper when the region is divided into 4 x 2 sections, suggesting that feature extraction methods that help to define the shape of the segmented region may help to improve the overall performance of the system. The results in the last row of Table 7 demonstrate the negative impact combining invariant moments and distance to centre features has on the overall classification performance.

Table 7. Classification of the percentage of area of a region divided into *n x m* sections.

Region division X x Y	MLP	NB	IB1	MB-AB	NBT	Dec T	AVG
2x1	71 (0.2)	70 (0.1)	61 (3.0)	71 (0.1)	71 (0.1)	71 (0.1)	**69**
2x2	70 (1.4)	71 (0.1)	63 (3.5)	71 (0.1)	71 (0.1)	71 (0.1)	**70**
3x2	76 (3.7)	76 (3.0)	68 (3.3)	72 (2.3)	77 (2.6)	76 (2.3)	**74**
3x3	76 (3.2)	76 (2.2)	69 (4.0)	73 (2.3)	75 (2.5)	74 (2.7)	**74**
4x2	83 (2.3)	81 (2.0)	81 (2.1)	77 (1.8)	82 (2.1)	80 (2.5)	**81**
4x3	75 (1.7)	75 (3.3)	71 (3.7)	72 (1.8)	74 (3.1)	76 (3.4)	**73**
4x4	74 (3.4)	74 (2.0)	71 (3.1)	72 (2.1)	76 (1.9)	76 (2.9)	**74**
4x2, DTC, Inv-M	78 (2.5)	73 (2.8)	74 (2.7)	74 (3.2)	77 (3.9)	80 (3.2)	**76**

Table 8 outlines the accuracy of the six classifiers on the 4x2 segmentation results in Table 7. After examining the classification rates for all the results presented in this paper (not included due to space restrictions) we have discovered that in more than 95% of cases the IB1 classifier produces a balanced error rate of no more than 2% difference between false positives and false negatives. From a purely quantitative assessment perspective this information, if proved sufficiently consistent, could provide us with a greater level of accuracy when assessing scallop abundance. However from an Artificial Intelligence perspective if is far more desirable to work towards reducing the error rate to a more acceptable level.

Table 8. Breakdown of the classification results for 4x2 segmentation in Table 7.

Segmentation 4x2	TP	TN	FP	FN
MLP	341	1022	145	140
NB	344	990	177	137
IB1	327	1009	158	154
Multi-AB	417	847	320	64
NB-Tree	314	1043	124	167
Dec-Table	313	1002	165	168

5 Conclusions and Further Work

A substantial amount of the research we have undertaken so far has been in developing a suitable multilayered system capable of automatically annotating

commercial scallop bed video footage. Our results so far show promise but also leave room for improvement.

The experimentation using only invariant moments demonstrate that this approach produces at best mediocre results. The results produced by dividing the segmented region into sections and measuring the area occupied by the region in each section indicate that the emphasis of further work should be on finding feature extraction techniques that better describe the overall shape of the segmented region. This work will include tests using sampling techniques and pattern classification using an n-tuple classifier system [7].

The research and results covered within this paper are based on still images extracted from the video footage. iNVT is also capable of working with multiple frames or video sequences as needed for integrating video footage into the process to provide an automated system. It will be necessary to develop a tracking system for *found* instances across consecutive frames to avoid counting scallops more than once.

We are currently working on a new dataset which will include a minimum of 500 scallop instances. This dataset will be thoroughly scrutinised for accuracy and will be used primarily for testing classifier learning performance. We are also looking at broadening the scope of the system to incorporate other underwater domains with similar characteristics. We have recently obtained video footage of seahorse activity and intend to test our system on this footage in the near future.

6 References

1. Wilson, A.: First steps towards autonomous recognition of Monterey Bay's most common mid-water organisms: Mining the ROV video database on behalf of the Automated Visual Event Detection (AVED) system. Technical Report, MBARI, CA, USA (2003).
2. Walther, D., Edgington, D.R., Koch, C.: Detection and Tracking of Objects in Underwater Video. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2004) 544-549
3. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition) Morgan Kaufmann (2005)
4. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. In: IEEE Transactions on Pattern Analysis and Machine Intelligence **11** (1998) 1254-1259
5. Itti, L.: The iLab Neuromorphic Vision C++ Toolkit: Free tools for the next generation of vision algorithms. In: The Neuromorphic Engineer (2004)
6. Gonzalez, R.C., Woods, R.E.: Digital Image Processing using MATLAB. Prentice Hall (2004)
7. Schechner, Y.Y., Karpel, N.: A Trainable n-tuple Pattern Classifier and its Application for Monitoring Fish Underwater. In: 7th International Conference on Image Processing and its Applications. Manchester, UK (1999) 255-259