Active learning for the classification of species in underwater images from a fixed observatory

Torben Möller
Biodata Mining Group, Bielefeld University
Bielefeld 33615, Germany
tmoeller@cebitec.uni-bielefeld.de

Ingunn Nilssen
Statoil ASA, Research and Technology
Trondheim 7005 Norway
innil@statoil.com

Tim W. Nattkemper
Biodata Mining Group, Bielefeld University
Bielefeld 33615, Germany
tim.nattkemper@uni-bielefeld.de

Abstract

Vision based wildlife monitoring is an important task in the field of environmental monitoring. Wildlife monitoring activities often create large collections of data needing computational approaches to (semi-) automated detection and annotation of objects in the images/video. In this work, we consider the special case of marine wildlife monitoring using camera equipped fixed observatories. In such cases where a-priori knowledge about which species to find is limited, a standard computer vision approach, employing supervised learning, will not be applicable for detecting and classifying species (or events) in the images.

In a recently proposed unsupervised learning method, image patches are extracted from a time series of underwater images that feature moving species (like starfish, etc). The patches are automatically grouped into clusters with similar morphology and a so called relevance score is assigned to each of the clusters describing the likeliness that it contains patches showing unusual changes. However, due to the unsupervised fashion (i) the categories don't have labels and (ii) do not reflect the species distribution satisfactory.

In this paper, we propose an active learning method that builds upon these results and can be used to assign taxonomic categories to single patches based on a set of human expert annotations making use of the cluster structure and relevance scores. The evaluation shows that compared to traditional sampling strategies our approach uses significantly less manual labels to train a classifier. We are confident that the results are relevant for non-marine contexts as well.

1. Introduction

In recent years, a growing number of so called *Fixed long-term Underwater Observatories* (FUO) [5, 2, 19, 9] equipped with fixed digital HD cameras have been deployed. These FUOs allow to monitor marine habitats over time, including long term changes in a reef [12] or the monitoring of particular species. Some effort has been spent on the (semi-) automatic classification of species for both terrestrial and marine applications. In these cases, often, a supervised classifier is trained. Examples are e. g. [3, 14] for terrestrial wildlife monitoring and [6, 15] for underwater wildlife monitoring. However, in environments with limited a-priori knowledge about the ecosystem it is often not practicable to train a classifier in a standard supervised fashion, as this approaches are based on a-priori knowledge about the species present. An unsupervised machine learning method for detection of short term changes in the visual field (e. g. occurrence of a specimen) in underwater images has been proposed in [11] for the Lofoten-Vesterålen (LoVe) FUO (Figure 1). Given a set of images, the method extracts a set of image patches, showing small regions where change occurred. The method, furthermore, groups all patches into clusters of similar change patterns and assigns a *relevance score* to every cluster describing how likely a patch shows an interesting change. In order to obtain a complete analysis of all species appearances for the entire observation period, the patches need to be assigned to taxonomic labels. The computationally detected and clustered patches must therefore be inspected and labeled by human experts (i. e. marine biologists). To speed up the manual labeling significantly, we propose an active learning approach in this paper.

Active learning is applied to facilitate the training of

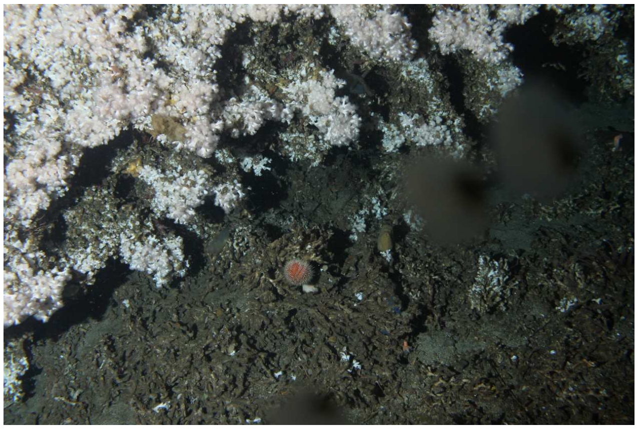

Figure 1. One sample image taken by the LoVe Ocean Observatory.

a supervised classifier to classify unsupervised clustered patches. The point of active learning is that training samples are chosen in a way that makes the classifier learn faster, i. e. the classifier needs less trainings samples to achieve a specific classification performance. Most active learning methods chose the training samples depending on the uncertainty of the classifier(s), the expected model change or the expected improvement of the classification performance. See [18] for an overview of the methods. In [13] an active learning method was proposed based on choosing training samples from clusters obtained using the k-medoids [8] algorithm. k-medoids is a clustering algorithm similar to k-means. The main difference is that the prototype of a cluster is not the mean but the medoid of the cluster, i. e. the element with the minimal distance to all elements in the cluster. When choosing training samples, the classifier considers not only the uncertainty of a sample but also gives priority to samples that are representatives of dense clusters.

The contribution of this paper is to propose a new active learning method for classification of underwater image patches (Section 2). Similar to the method in [13], the proposed method makes use of a prior clustering of the samples. In contrast to other methods, a *relevance score* (computed as described in [11]) assigned to the clusters is used to further improve the selection of the training samples. The success of the strategy of selecting the training samples is illustrated in Section 4. Section 5 discusses the method and its evaluation.

2. Methods

In active learning (like in other supervised machine learning), a classifier is trained with a labeled set of trainings samples. In contrast to other supervised machine learning scenarios, in an active learning scenario, the learner actively decides which samples have to be labeled and added to the training set. An active learning scenario for classification (see Figure 2) is described by (i) a sampling strategy for choosing unlabeled data (ii) an oracle (often a human expert) for labeling the chosen samples and (iii) a classifier (e. g. Bayes classifier, SVM) for automatic classification of the unlabeled data. Given these three components, the active learner works as follows: First, from a pool of unlabeled samples, one sample is selected according to the sampling strategy. Next, the chosen sample is labeled by the oracle and added to the training set. The training set is used to fit the classifier to the data and the process is repeated. In the end, the trained classifier is used to classify the complete dataset.

In contrast to the majority of existing strategies, we do not involve the classifier in the sample selection process. The complete trainings set is chosen according to the proposed strategy (see below) before the labels are determined (see Figure 2 dashed arrows). In the present case, the labels are determined by a human expert. Thus, it remains to describe the sampling strategy and the classification in the rest of this section.

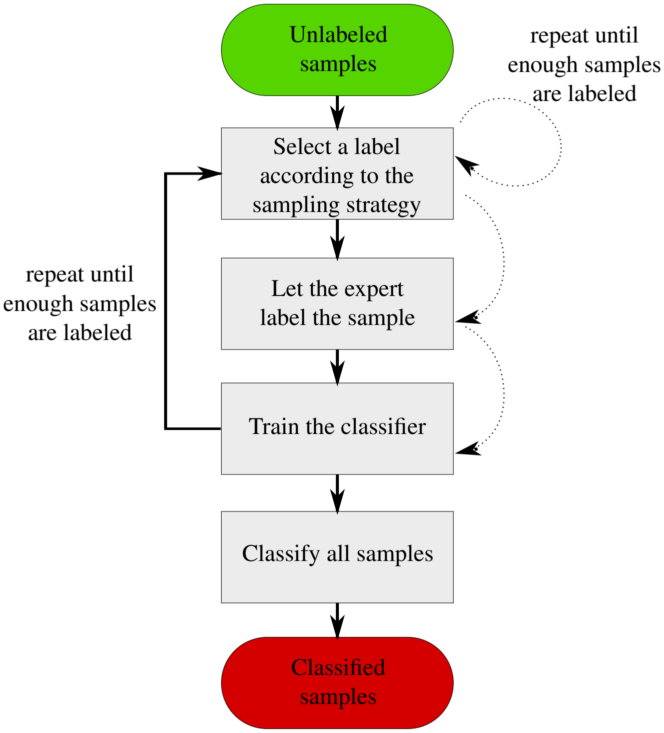
Unlabeled samples
repeat until enough samples are labeled
Select a label according to the sampling strategy
repeat until enough samples are labeled
Let the expert label the sample
Train the classifier
Classify all samples
Classified samples

Figure 2. In a typical active learning scenario, the steps ‘select a sample’, ‘label the sample’, ‘train the classifier’ are repeated in cycles. If the sampling strategy does not involve the state of the classifier, it is possible to first select all training samples and then label the samples (dashed arrows).

Table 1. The most frequently used symbols

Notation	Meaning
F_j	the feature vector of the j-th sample
L	the number of labels
ℓ	an integer representing a label
c_j	the cluster index of the sample j
r_i	the relevance score of cluster i
δ_j	the indicator function of the labeled samples
ℓ_j	The true label of sample j
$\hat{\ell}_i$	The cluster label
$\tilde{\ell}_j$	The label predicted for sample j

2.1. Notation and setup

Let N be the number of samples (i. e. image patches) $s_1, \ldots, s_N$ to be classified. We assume that a feature representation of the samples is given as feature vectors $F_1, \ldots, F_N$. Moreover, we assume that a set of integers l (called *labels*) is given representing semantic classes (e. g. : ‘starfish’). The number of semantic classes is denoted by L. Without loss of generality let $1 \leq l \leq L$ for each label l.

The proposed method expects that the features are already clustered into M clusters and that a *relevance score* is assigned to each cluster. The clustering can be done by any clustering method, including the method proposed in [11]. The index of the cluster containing the sample j will be denoted by $c_j \in \{1, \ldots, M\}$. The relevance scores should describe how likely patches in a cluster show foreground (i. e. moving species). Clusters can be determined as described in [11]. The relevance score of the cluster with index i will be denoted by r_i. For a quick reference of the most frequently used notations see Table 1.

2.2. Sampling

The automated sampling strategy has to balance the following two almost conflicting demands:

1. Samples from a cluster with a high relevance score, are more likely to be selected than samples from a cluster with a low relevance score.
2. Each sample (whichever cluster it belongs to) has a chance to be selected.

To implement these demands, the strategy is to first select a cluster deterministically and then draw a sample from the cluster randomly. At any iteration, let f_i denote the number of times, a patch from cluster i has been drawn. To any cluster i we assign a score balancing the relevance and the accumulative activity of cluster i.

$$h_i = \begin{cases} \infty & \text{if } f_i = 0 \\ r_i/f_i & \text{else} \end{cases} \tag{1}$$

The cluster i' to draw the next sample from is then given by

$$i' = \underset{1 \leq i \leq M}{\arg\max}\,(h_i). \tag{2}$$

For a sample s_j $(1 \leq j \leq N)$ let

$$\delta_j = \begin{cases} 1 & \text{the sample } j \text{ has been annotated} \\ 0 & \text{else} \end{cases} \tag{3}$$

denote the indicator function of the set of labeled samples. The next sample is then drawn from the set

$$\{s_j | c_j = i' \wedge \delta_j = 0\}. \tag{4}$$

As we assume the expert to always label correctly, the label assigned to the sample j by the human expert is considered the ground-truth label ℓ_j.

2.3. Classification

For classification, a Support Vector Machine (SVM) [4] with a radial basis kernel [17] is used. Multi-class classification is realized with the one-vs-one method [10]. The SVM is trained with all samples. To do so, we propagate the labels from the human expert as follows: First, the expert labels are propagated to the clusters. For each cluster i, we define the cluster label by majority voting:

$$\hat{\ell}_i = \underset{1 \leq l \leq L}{\arg\max} \left| \{1 \leq j \leq N \,| \delta_j = 1 \wedge c_j = i \wedge \ell_j = l\} \right|. \tag{5}$$

Next, the labels of the clusters are propagated to the samples. The label of a sample j for the SVM training is defined by

$$\hat{\ell}_{c_j} \tag{6}$$

i. e. the label of the cluster containing j.

The training set used to fit the the SVM is then given by the (sample, label)-pairs

$$\left\{ (F_j, \hat{\ell}_{c_j}) | 1 \leq j \leq N \right\}. \tag{7}$$

The classifier finally is used to classify all samples. The label predicted for the sample j by the SVM classifier is denoted by $f(F_j)$. As we assume the expert to always label the samples correctly, the final label of a sample s_j is given by

$$\tilde{\ell}_j = \begin{cases} \ell_j & \delta_j = 1 \\ f(F_j) & \text{else} \end{cases} \tag{8}$$

The effect on the evaluation caused by using the labels $\tilde{\ell}_l$ instead of the classifier outputs $f(F_l)$ will be discussed in Section 5.

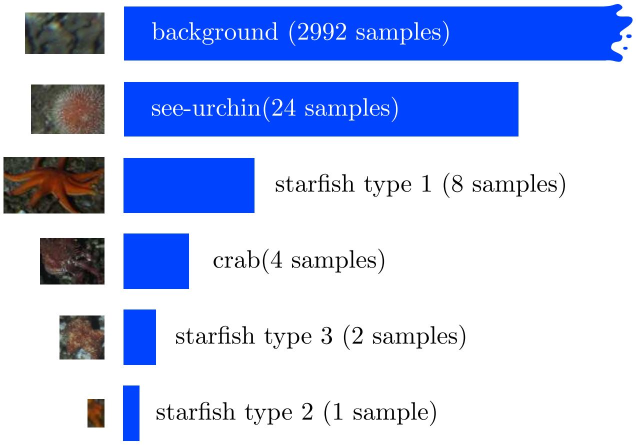

Figure 3. The figure shows the class frequencies and one example of each class. Note that the bar depicting the frequency of background patches is not true to scale for means of better visualization.

3. Material

The proposed method is evaluated on underwater images provided by the Lofoten Vesterålen (LoVe) Ocean Observatory [5]. LoVe is a fixed long-term underwater observatory monitoring a coral reef at (N 68° 54.474′, E 15° 23.145′) in the Norwegian Sea. The reef is located 22 km off the coast in a water depth of approximatively 260 m.
The observatory was deployed in October 2013 and takes one image every 60 minutes. All images are publicly available online at `http://love.statoil.com/`. A sample image of the LoVe images is shown in Figure 1.

The change detection method proposed in [11] has been applied to a subset of 24 images taken by the LoVe ocean observatory. The change detection method extracted 3031 image patches divided into 6 categories: Background, Crab, Starfish type 1, Starfish type 2, See-urchin and Starfish type 3. As can be seen in the histogram in Figure 3, the dataset is imbalanced as most of the patches show background.

4. Evaluation

For evaluation, we applied the proposed method to the image patches described in the previous section (see Figure 3). The cluster indices c_j from the previous clustering and the relevance scores r_i for the classes are given by the method in [11]. The features Fj are extracted using *dominant color*. To do so, the colors in a patch are quantized using the modified median cut algorithm [1], i. e. a color palette C of 5 colors is generated. The color of each pixel p_k in the patch is assigned to the nearest color γ_k in the palette. The *dominant color* is then $\arg\max_{\gamma \in C} |\{p_k | \gamma_k = \gamma\}|$, where the p_k denote the pixels in the patch.

The evaluation of the method is twofold: First, the method is compared to other methods by testing how quick the learner learns from the drawn samples. Second, the overall method is evaluated by testing its performance after a reasonable number of samples has been labeled manually.

4.1. Comparison to other methods

To evaluate the sampling rule, we follow a common strategy: For each number n ($1 \leq n \leq N$) a training set of size n is selected according to the proposed sampling strategy and the classifier is trained with the selected samples. Each time, a performance measure is computed for the classifier. The progression of the performance measure with increasing number of training samples reflects the quality of the sampling strategy. A performance measure often used for this type of evaluation is the accuracy, defined as the percentage of correct predictions. However, the accuracy is a poor measure if the dataset is imbalanced as it is often the case in real-world scenarios (including the dataset used here, see Figure 3) especially in wildlife monitoring. A naive classifier that maps everything to the most abundant class, would always have a *good* accuracy although it misclassifies all samples from the other (potentially more important) classes (compare [16]). For an appropriate evaluation of the method, we use a slightly modified version of accuracy that does not involve correctly classified background. For this, we assume without loss of generality that the class *background* has the label 0. Moreover, we define

$$\widehat{\mathrm{TP}} = \left| \left\{ 1 \leq j \leq N \middle| \ell_j \neq 0 \wedge \ell_j = \tilde{\ell}_j \right\} \right| \tag{9}$$

to be the number of correctly classified foreground samples and

$$\widehat{N} = \left| \left\{ 1 \leq j \leq N \middle| \tilde{\ell}_j \neq 0 \vee \ell_j \neq 0 \right\} \right| \tag{10}$$

to be the number of samples being either foreground or classified as foreground. The *foreground accuracy* is then defined by

$$\widehat{a} = \frac{\widehat{\mathrm{TP}}}{\widehat{N}}. \tag{11}$$

We denote by $\widehat{a}(n)$ the accuracy that has been measured for a classifier trained with n samples.

We compare our method to (i) random sampling and (ii) uncertainty sampling. For each, we use a modification of the approach in Section 2 that makes neither use of (a) the precomputed clusters nor (b) the proposed sampling strategy based on the cluster relevances. The first modification (according to (a)) is that the SVM is not trained with the propagated labels (see Equation 7) but only with the set

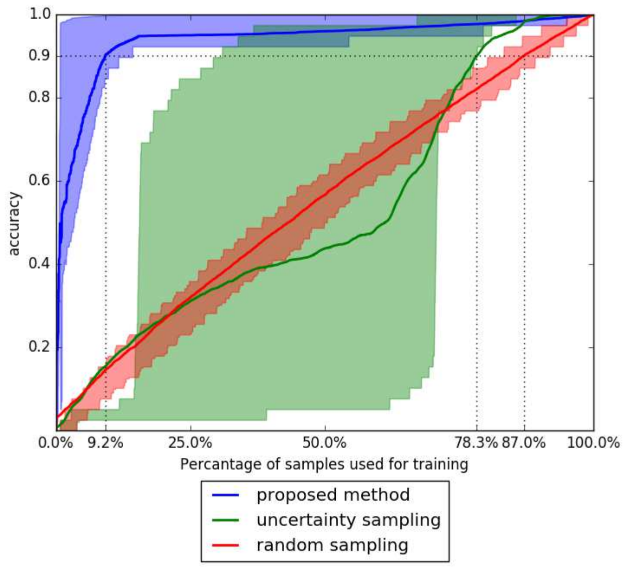

Figure 4. The figure compares the performance of the proposed sampling strategy with the performance of random sampling. The colored areas illustrate the range between the 0.25 percentile and the 0.75 percentile. The black dotted lines indicate how many samples have to be labeled before a method achieves at least a 90% accuracy. The proposed method performs better than uncertainty sampling and random sampling. It can also be seen that uncertainty sampling performs worse than random sampling when about 20% to 70% of the samples are labeled. See Section 5 for a discussion of this phenomenon.

$\{(F_j, \ell_j) | 1 \leq j \leq N \wedge \delta_j = 1\}$ of expert labels. The second modification (according to (b)) is that we do not use the sampling strategy proposed in Section 2.2, but draw

(i) randomly with uniform distribution from all samples not labeled by the expert.

(ii) according to uncertainty sampling as described in [18] as soon as samples from at least two different classes have been labeled by the expert. According to random sampling when all expert labels are from one class.

We ran each algorithm 200 times generating accuracies $\widehat{a}_k^{\text{rel}}(n)$ (for the proposed sampling), $\widehat{a}_k^{\text{rnd}}(n)$ (for random sampling) and $\widehat{a}_k^{\text{unc}}(n)$ (for uncertainty sampling) for $1 \leq k \leq 200$ and $1 \leq n \leq N$. The average accuracies for the proposed method, random sampling and uncertainty sampling are defined by

$$\widehat{a}^{\text{rel}}(n) = \frac{1}{200} \cdot \sum_{k=1}^{200} \widehat{a}_k^{\text{rel}}(n) \quad (1 \leq n \leq N) \quad (12)$$

$$\widehat{a}^{\text{rnd}}(n) = \frac{1}{200} \cdot \sum_{k=1}^{200} \widehat{a}_k^{\text{rnd}}(n) \quad (1 \leq n \leq N) \quad (13)$$

$$\widehat{a}^{\text{unc}}(n) = \frac{1}{200} \cdot \sum_{k=1}^{200} \widehat{a}_k^{\text{unc}}(n) \quad (1 \leq n \leq N) \quad (14)$$

respectively.
Figure 4 shows the accuracies obtained using (i) the proposed sampling strategy (ii) uncertainty sampling and (iii) random sampling. It can be seen that the proposed method learns much faster than the other methods. As indicated by the black dotted line, random sampling needs 87% (2638 labels) of the samples to be labeled to obtain a 90% foreground accuracy and the uncertainty sampling needs $78,3\%$ (2372 samples) of the samples while the proposed method needs only 9.2% (279 labels) of the samples to be labeled.

4.2. Evaluation of the method

To evaluate the proposed method, we analyze the results after training the classifier with 300 samples selected according to the proposed sampling strategy. In contrast to the previous subsection, the experiment has only been conducted once. Figure 5 gives an overview of the classification. We compute the most common performance measures separately for every class: Given a label l, we define by

$$\text{P}_l = |\{1 \leq j \leq N \,|\ell_j = l\}| \quad (15)$$

$$\text{PP}_l = \left|\left\{1 \leq j \leq N \,\middle|\tilde{\ell}_j = l\right\}\right| \quad (16)$$

$$\text{TP}_l = \left|\left\{1 \leq j \leq N \,\middle|\tilde{\ell}_j = l \wedge \ell_j = l\right\}\right| \quad (17)$$

the number of positives, predicted positives and true positives, respectively. Precision, recall and F1-score for class l are then defined by

$$\text{precision}_l = \frac{\text{TP}_l}{\text{PP}_l}, \quad \text{recall}_l = \frac{\text{TP}_l}{\text{P}_l} \quad \text{and} \quad (18)$$

$$\text{F1-Score}_l = 2 \cdot \frac{\text{precision}_l \cdot \text{recall}_l}{\text{precision}_l + \text{recall}_l}. \quad (19)$$

From Table 2 it can be seen that the average F1-score is

$$\overline{\text{F1-score}} = \frac{1}{L} \sum_{l=1}^{L} \text{F1-Score}_l = 0.93 \quad (20)$$

Moreover, it can be seen that the method performs well for all classes except for the crabs. Possible reasons for this are discussed in the next section.

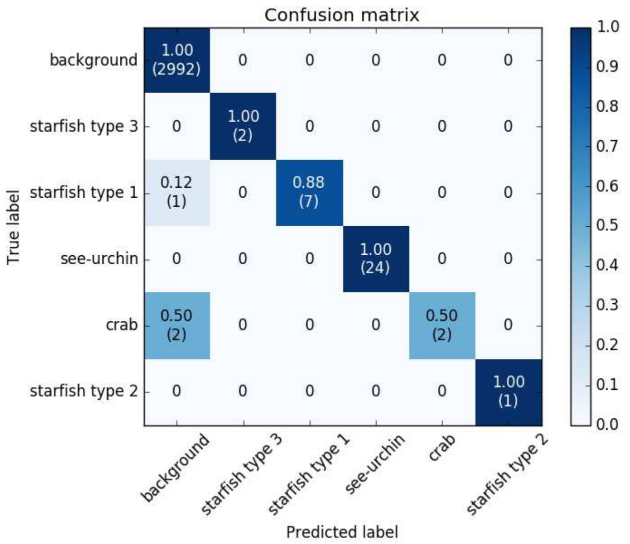

Figure 5. Confusion matrix after training the classifier with 300 labeled samples. The number in brackets at (i, j) is the total number of samples with true label i and predicted label j. The fractional number at (i, j) is the fraction of samples with predicted label j of all samples with true label i. Consequently, the recall of a class is found in the corresponding element on the diagonal.

Table 2. The precision, recall and F1-score for each class

class	precision	recall	F1-score
background	1.00	1.00	1.00
starfish type 3	1.00	1.00	1.00
starfish type 1	1.00	0.88	0.93
see-urchin	1.00	1.00	1.00
crab	1.00	0.50	0.67
starfish type 2	1.00	1.00	1.00

5. Discussion and Conclusion

In this work, we have addressed the problem of annotating image regions showing mobile species in a marine wildlife monitoring scenario. We have shown, how active learning can be combined with an unsupervised object detection framework to automatically classify large numbers of objects without an extensive labeling of many training samples. We compared our approach with two baseline methods: random sampling and uncertainty sampling. In both cases we could show that our approach needed much less training samples, i. e. learned much faster than the baseline and thereby also requires less time to perform the manual labeling.

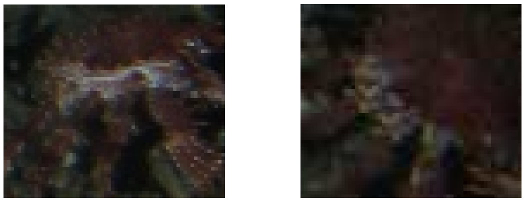

Figure 6. 'Crab' samples that have been classified as background.

The evaluation of the classification showed that the method performs good for five out of six classes. However, only two of 4 samples from the class 'Crab' have been found (see Figure 6). A possible reason is that the features of the 'crab' class are not well enough separable from the features of the 'background' class. This suggests that features have to be selected carefully for the image set and the monitored habitat.

It is a widely accepted practice that the trainings set used to fit a classifier and the test set used to test the classifiers performance are disjoint. A criticism regarding the evaluation can be that (in contrast to this practice) the test set was a subset of the trainings set in our evaluation. However, we believe that in the special case of this active learning application it is appropriate to include the training set in the test set since (i) in an practical application, the trainings set is always a part of the patches that have to be classified and (ii) the selection of an appropriate trainings set is an important part of the method and should be considered in the evaluation.

Figure 4 shows that uncertainty sampling performed worse than random sampling when the classifier was trained with about 20% to 70% of the samples. This might seem unintuitive at first glance, but the reason for this is that the number of classes is not known a-priori. Uncertainty sampling prefers to select samples near the *border* of two classes. A sample of an unknown class (i. e. no sample of that class has been labeled by the expert) is less likely near the border than a sample from of the classes that define this border. Thus, some classes remain unknown longer (i. e. until more samples have been labeled) than when random sampling is used and this causes the poor accuracy.

Acknowledgment

Financial support was given by Statoil ASA, Research and Technology, Norway

References

[1] D. Bloomberg. color quantization using modified median cut.

[2] O. N. Canada. NEPTUNE in the NE Pacific, 2014.

[3] G. Chen, T. X. Han, Z. He, R. Kays, and T. Forrester. Deep convolutional neural network based species recognition for wild animal monitoring. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 858–862, Oct 2014.

[4] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[5] O. R. Godø, S. Johnson, and T. Torkelsen. The love ocean observatory is in operation. *MARINE TECHNOLOGY SOCIETY JOURNAL*, 48, 2014.

[6] P. X. Huang, B. J. Boom, and R. B. Fisher. Hierarchical classification with reject option for live fish recognition. *Machine Vision and Applications*, 26(1):89–102, Jan 2015.

[7] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[8] L. Kaufman and P. Rousseeuw. Clustering by means of medoids. *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pages 405–416, 1987.

[9] K. Kawabata, F. Takemura, T. Suzuki, K. Sawai, E. Kuraya, S. Takahashi, H. Yamashiro, N. Isomura, and J. Xue. Underwater image gathering by utilizing stationary and movable sensor nodes: Towards observation of symbiosis system in the coral reef of okinawa. *International Journal of Distributed Sensor Networks*, 10(7), 2014.

[10] S. Knerr, L. Personnaz, and G. Dreyfus. *Single-layer learning revisited: a stepwise procedure for building and training a neural network*, pages 41–50. Springer Berlin Heidelberg, Berlin, Heidelberg, 1990.

[11] T. Möller, I. Nilssen, and T. W. Nattkemper. Change detection in marine observatory image streams using bi-domain feature clustering. In *International Conference on Pattern Recognition (ICPR 2016)*, 2016.

[12] T. Möller, I. Nilssen, and T. W. Nattkemper. Data-driven long term change analysis in marine observatory image streams. In *2016 ICPR 2nd Workshop on Computer Vision for Analysis of Underwater Imagery (CVAUI)*, pages 13–18, Dec 2016.

[13] H. T. Nguyen and A. Smeulders. Active learning using preclustering. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 79–, New York, NY, USA, 2004. ACM.

[14] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, C. Packer, and J. Clune. Automatically identifying wild animals in camera trap images with deep learning. *CoRR*, abs/1703.05830, 2017.

[15] J. Osterloff, I. Nilssen, and T. W. Nattkemper. A computer vision approach for monitoring the spatial and temporal shrimp distribution at the LoVe observatory. *Methods in Oceanography*, 2016.

[16] F. J. Provost, T. Fawcett, et al. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In *KDD*, volume 97, pages 43–48, 1997.

[17] B. Scholkopf, K.-K. Sung, C. J. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE transactions on Signal Processing*, 45(11):2758–2765, 1997.

[18] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

[19] Vardaro et al. A Southeast Atlantic deep-ocean observatory: first experiences and results. *Limnology and Oceanography: Methods*, 11:304–315, 2013.