Image Analysis Techniques to Accompany a new In Situ Ichthyoplankton Imaging System

G. Tsechpenakis, IEEE member, C. Guigand, and R.K. Cowen

Abstract— We have built a high resolution towed digital imaging system (ISIIS) capable of imaging water volumes sufficient to accurately quantify even rare plankton (e.g. larval fish) in situ. This imaging system produces very high resolution imagery at very high data rates necessitating automated image analysis. As we are interested in the identification and quantification of a large number of organisms, some of which are relatively similar to each other, we are developing an automated system for detection and recognition of organisms of interest using computer vision tools. Our method aims at (i) the detection of multiple regions (organisms) of interest automatically, while filtering out noise and out-of-focus organisms, and (ii) the classification of the detected organisms into pre-defined categories using shape and texture information. For the organisms detection, we use a probabilistic scheme based on image statistics to locate the regions of interest and – based on size and shape constraints – we filter out the noise, i.e., regions that are detected but do not correspond to organisms. For the classification of the detected organisms, we use the Scale Invariant Feature Transform (SIFT) for matching between the detected regions and the organism images in our database (“dictionary”).

Index Terms—Automated analysis for ISIIS, computer vision, detection of organisms in plankton images, organism recognition.

I. INTRODUCTION

Current technologies available for the study of plankton remain limited in comparison to the spatial-temporal resolution and data acquisition rate available for physical oceanographic measurements.

Specifically, plankton measurements are made primarily by use of net collections, versus high speed digital output possible for physical sampling. Though net technology has become quite sophisticated (e.g. MOCNESS), enabling vertical resolution coupled with detailed physical data, net samples still require the task of being processed manually, which is a time-consuming and costly effort. Further, nets integrate organisms over the sampling distance and depth, significantly reducing sample resolution.

To address this problem, we have built a high resolution towed digital imaging system (ISIIS) capable of imaging water volumes sufficient to accurately quantify even rare plankton (e.g. larval fish) in situ (Cowen and Guigand, unpubl. data). This imaging system produces very high resolution imagery at very high data rates necessitating automated image analysis. As we are interested in the identification and quantification of a large number of organisms, some of which are relatively similar to each other, we are developing an automated system for detection and recognition of organisms of interest using Computer Vision tools.

The existing approaches on similar problems either (i) assume that the organisms of interest are already segmented and they focus on the classification problem, or (ii) focus on shape features to be used in the classification scheme, without taking into account texture information.

The goal of our system is to use as input the high resolution images and to: (i) detect the multiple regions (organisms) of interest automatically, while filtering out noise and out-of-focus organisms, and (ii) classify the detected organisms into pre-defined categories using shape and texture information. The organism detection scheme uses a background subtraction approach based on image statistics, namely the grayscale intensity distribution of the expected background; then, based on size and shape constraints we filter out the noise, i.e., regions that are detected but do not correspond to organisms. The main advantage of this detection approach is that it is probabilistic, takes into account possible local variations in the background intensity, and it provides higher robustness compared to common intensity thresholding. For the recognition scheme we use prior knowledge about the shape and texture of the expected organisms (i.e., based on an image library). We model the texture and shape of the organisms with the Scale Invariant Feature Transform (SIFT), commonly used in Computer Vision applications, e.g., for image and shape registration.

In our results we show the performance of our method for the detection and recognition of organisms with qualitative (ISIIS images) and quantitative (Receiver Operating Characteristic [ROC] curves) examples. Our ultimate goal is to make the system real-time to assist in adaptive sampling of oceanographic features.
II. ORGANISMS DETECTION IN PLANKTON

A. Existing Approaches for detection

The detection/segmentation of regions of interest (ROIs) has been one of the main topics in Computer Vision in the last couple of decades. A popular general approach to this problem is the background modeling \cite{3, 2, 10, 9, 8}, so that the target regions are detected with background subtraction. This approach requires non-parametric modeling of the background color, grayscale intensity or texture, and its robustness depends on how effective this modeling/approximation is. One way to approximate the background color or intensity distribution sufficiently is using a mixture of Gaussians (mixture of parametric distributions) \cite{3, 9} or a non-parametric kernel density estimator \cite{2}. An extensive survey on background estimation and a combined method is presented in \cite{10}.

B. Our detection approach

Let $\Omega$ be the image domain, $B \subseteq \Omega$ be the background region, and $\Omega \setminus B$ be the region of interest (Fig. 1(a)). To model the background grayscale intensity distribution, we estimate the probability density function $p_B$ of the intensity values obtained from sample (training) background regions. Since a single parametric distribution (e.g. Gaussian) cannot sufficiently model the background, we use a mixture of Gaussians,

$$p_B = \sum_{i=1}^{n} a_i \cdot g(\mu_i, \sigma^2_i),$$

where the Gaussian parameters $(\mu_i, \sigma^2_i)$ and the weights $a_i$ are estimated by the Expectation Maximization (EM) algorithm \cite{1}. For our application we used $n=3$ Gaussians, since it is experimentally shown that 3-4 Gaussians are sufficient to model any nonparametric distribution. Fig. 1(b) illustrates the three Gaussians (red, blue, green), and their weighted combination (black), as estimated by the EM algorithm using training samples from the background.

Then, the probability of an image pixel $x_i \in \Omega$ with grayscale intensity value $I(x_i)$ being consistent with (belonging to) the background is,

$$P(x_i | B) = \frac{I(x_i | B) \cdot dI}{\int_{I(x_i | B) \cdot dI}},$$

where $dI$ determines a small grayscale interval.

After estimating the probability map of the background, we can estimate the probabilities of the image pixels belonging to the regions of interest (Fig 2) as $P(x_i | \Omega \setminus B) = 1 - P(x_i | B)$. Fig.3(b) illustrates the (locally) thresholded probability map of the regions of interest of the example of Fig. 2(a): the black regions/pixels correspond to the detected regions of interest (ground-truth with red boxes), but there are some inaccuracies due to clutter and out-of-focus objects.

To eliminate the noisy effects in the probability map, and apart from the size constraint that we impose, we smooth the probability map using a discriminative Conditional Random Field (CRF) \cite{4}; after thresholding the new probability field locally, we eliminate significant amount of noise, as shown in Fig. 2(c).

The formulation of the discriminative CRF is as follows. Let $L = \{l_i\}$ be the labels associated to the image pixels; in our case, a label can have two values, i.e., a pixel belongs to the region of interest (black) or the background (white) (Fig. 2(b)). The discriminative CRF can estimate directly the labels distribution given an appropriate set of image features as,

$$p(L | I) = \frac{1}{Z} \exp \left\{ \sum_{i \in \Omega} \psi_a(l_i, I(x_i)) + \sum_{i \in \Omega} \sum_{j \in \Omega} \psi_i(l_i, l_j, I(x_i), I(x_j)) \right\},$$

where $Z$ is a normalization constant, $||\Omega||$ is the size of the image domain, and $N_i$ indicates the neighborhood of the $i$-th pixel. In this definition, $\psi_a$ is called association potential, since it associates the label of a pixel with its intensity. Also, $\psi_i$ is called interaction potential, since it determines the interaction between the neighboring pixels, in terms of both their intensity values and their labels. Note that common Random Fields can be expressed in terms of an association and interaction potential, with the difference that usually only the interactions between labels are taken into account in the interaction potential (which causes the known label bias problem \cite{5}).

The association potential of eq. (3) is defined as the log likelihood,

$$\psi_a(l_i, I(x_i)) = \log P(l_i | I(x_i)),\quad (4)$$

whereas the interaction potential between two neighboring pixels is given by,
where \( z_{in} \) is a normalization constant, and \( I(x_i), I(x_j) \) are the grayscale intensity values of at the pixel locations \( x_i \) and \( x_j \) respectively. The interaction potential is responsible for smoothing the probability field defined in eq. (3), if two neighboring pixels have different intensity values; otherwise the interaction potential has no effect.

One can see in Fig. 2(c) that the labeled image (black for the regions of interest and white for the background) still contains noisy pixels and/or regions. Also, after smoothing the probability field, we lose some “informative” pixels, i.e., parts of the regions of interest. Fig. 3 illustrates the detection results (red boxes) superimposed on the original image, along with the ground-truth (numbered blue boxes); comparing the ground-truth with the resulting regions, one can see that there is one region (blue box (5)) that was not detected, while the remaining six regions were detected successfully. Also, ground-truth regions (1), (2), (4), (6) and (7) where detected successfully, i.e., the entire object of interest is inside the extracted region, for the case of region (3) only a part of the organism is detected. To address this problem, as well as to eliminate the false positives (detected regions that do not correspond to an object of interest), we need to have the robust recognition scheme that we describe below, which can deal with missing data and can efficiently exclude these false positives.

III. SIFT-BASED RECOGNITION OF THE ORGANISMS

In order to describe the shape and texture of the extracted objects of interest, and to use this description for recognition, we use the Scale Invariant Feature Transform (SIFT). SIFT descriptors were introduced by D. Lowe [7,6] and they are commonly used today in Computer Vision applications, mainly for image/shape retrieval and registration.

A. The Scale Invariant Feature Transform (SIFT)

In contrast to most of the existing approaches for local feature generation, this method transforms an image into a collection of local feature vectors, each one of which is invariant to image translation, scaling, and rotation, and invariant to illumination changes and affine or 3D projection. Therefore, SIFT provides robustness to small changes in the objects appearance in the examined images and the recognition phase is more efficient. The scale-invariant features are efficiently identified by using a staged filtering approach:

1) Scale-space extrema detection: The first stage of key-point detection is to identify locations and scales that can be repeatably assigned under differing views of the same object. Detecting locations that are invariant to scale changes of the image can be accomplished by searching for stable features across all possible scales, using a continuous function of scale known as scale-space. The scale-space of an image is defined as a function \( L(x, \sigma) \) that is produced from the convolution of a variable-scale Gaussian \( g(x, \sigma) \) with the input image \( I(x) \):

\[
L(x, \sigma) = g(x, \sigma) \ast I(x),
\]

(6)

where \( \sigma \) and \( x \) indicate the scales and the image locations respectively. To detect stable key-point locations in scale-space, the scale-space extrema are detected in the difference-of-Gaussians function convolved with the image, which can be computed from the difference of two nearby scales separated by a constant factor \( k \):

\[
D(x, \sigma) = [g(x, k \sigma) - g(x, \sigma)] \ast I(x) = L(x, k \sigma) - L(x, \sigma)
\]

(7)
In order to detect the local maxima and minima of $D(x, \sigma)$, each point is compared to its neighbors in the image domain and its neighbors in the scale domain (above and below the current scale). A point is selected as local extremum only if the value of $D$ at this position is larger or smaller than the values of $D$ at all of the neighbors. The cost of this check remains low since most points are eliminated after the first few checks.

2) Key-point localization: Once a key-point candidate has been found, the next step is to perform a detailed fit to the nearby data for location, scale, and ratio of principal curvatures [6]. This information allows points to be rejected that have low contrast and thus are sensitive to noise.

3) Orientation assignment: One or more orientations are assigned to each key-point location based on local image gradient directions. All operations are performed on image data that has been transformed relative to the assigned orientation, scale, and location for each feature, providing invariance to these transformations. Then, an orientation histogram is estimated from the gradient orientations of sample points within a region around each key-point. Each sample added to the histogram is weighted by its gradient magnitude and by a Gaussian-weighted circular window. Peaks in the orientation histogram correspond to dominant directions of local gradients. The highest peak in the histogram is detected, and then any other local peak that is within a percentage (usually 80%) of the highest peak is used to also create a key-point with that orientation. Therefore, for locations with multiple peaks of similar magnitude, there will be multiple key-points created at the same location and scale but different orientations.

4) Key-point descriptor: the local image gradients are measured at the selected scale in the region around each key-point. These are transformed into a representation that allows for significant levels of local shape distortion and change in illumination.

B. SIFT-based recognition

After estimating the key-points of the extracted regions, we use the sample specimens from our library for recognition: we detect the number of matches between the input key-points and the corresponding points of the samples. In case the library contains a large number of samples for a wide variety of specimens to be recognized, one could follow two alternative approaches using the SIFT descriptors:

1) The use of Principal Component Analysis (PCA) on the extracted key-points and matching in the lower-dimensional space. The advantage of PCA is that it reduces the recognition complexity by reducing the dimensionality of the feature space; on the other hand, dimensionality reduction leads to less accurate recognition results.

2) The use of a boosting-alike supervised learning (e.g., AdaBoost), which avoids the time-complexity of one-by-one matching. This approach reduces the recognition complexity by learning off-line the specimens’ feature manifolds; the main drawback of this approach in our application is the large number of outliers, i.e., the large number of correctly detected regions that are recognized with high uncertainty.

<table>
<thead>
<tr>
<th>Specimen</th>
<th>Number of sample images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trichodesnium</td>
<td>248</td>
</tr>
<tr>
<td>Larvaeans</td>
<td>221</td>
</tr>
<tr>
<td>Fish larvae</td>
<td>203</td>
</tr>
<tr>
<td>Copepods</td>
<td>227</td>
</tr>
<tr>
<td>Chaetognaths</td>
<td>212</td>
</tr>
</tbody>
</table>

At this stage, as we describe in the following section, our library contains five specimens and in total 1,110 sample images. Since the low complexity of the recognition is beyond our current goals, and the size of our library is limited, we follow one-by-one matching between the input (detected) regions and the library samples.

IV. Experimental results

We built a library containing images (manually cropped regions) of five taxa of interest, namely Trichodesnium, Larvaeans, Fish larvae, Copepods, and Chaetognaths. Table I shows the distribution of the samples per taxon. This is the first version of a general library that we want to build, which will contain a wide variety of cropped specimen images, along with their SIFT descriptors.

In this section we show our results in three case studies: (i) the estimation of the best matching sample for a predetermined specimen, (ii) the recognition of an input organism using all library samples (from all five sample taxa), and (iii) the recognition of all detected regions of interest.

A. Case study #1: matching within a pre-specified category

Fig. 4(a) illustrates an example of a detected specimen to be recognized. The specimen was manually identified as Trichodesnium and it was detected in the region that corresponds to the ground-truth region (7) of Fig. 3. In our library we included samples of this taxon and we used those samples for SIFT-based matching with the input region (a); Fig. 4(b) illustrates some of the samples of the specific specimen. In this figure, one can see that the samples are in different scales and orientations, and their appearance is distorted (deformed) due to the off-plane rotated position of the specimen. In Fig. 4(c) we show the matching between the key-points of the input region and the best-matching sample: both regions were automatically scaled and 6 matches were detected; the correspondences between the matched key-points are illustrated in red lines. Table II shows the key-points estimated for all nine samples, along with the matched points between the input region and these samples. The maximum number of detected “matches” (matched points) indicates the best matching sample.

B. Case study #2: recognition of a single detected region

In our second case-study experiment, we extracted the SIFT
The two leftmost columns of Table III correspond to indicative samples from our library, along with their taxon labels; in this table we illustrate two samples from each category. Similarly to the previous experiment, we estimated the SIFT key-points of the input region, as detected in the original input image (Fig. 3, ground-truth box (7)). Then we performed one-by-one matching between the input and sample key-points. Table III illustrates the number of the detected key-points for each image (third column), along with the matching results (rightmost column). Again, the best matching sample is the one that gives the maximum number of “matches”, i.e., the maximum number of key-points found in both images (input and sample). Then, the label of the detected region is the label of the best matching sample, which in this experiment was *Trichodesmium*.

### Case study #3: recognition of multiple detected regions

To obtain a first evaluation of the overall recognition performance, we estimated the ROC (Receiver Operating Characteristic) curves for the recognition of all detected specimens in an input image. In Table IV we show the ground-truth for the 10 input images. In the testing phase, we first detected the regions of interest and then performed the SIFT-based matching.

In Fig. 5 we illustrate the ROC curve (sensitivity vs. specificity function) for each specimen in our library, along with the recognition rate. One can see that we achieve the highest recognition rate for the *Chaetognath* taxon. The lower recognition rates for the rest of the taxa are mainly due to (i) the out-of-focus appearance of the regions, and (ii) the small size of the regions that lead to detection failure. Note that it is not necessary to use ROC curves to evaluate the performance of the organism detection in the original frame separately from the recognition stage, since any detection failures (either false positives or false negatives) directly affect the recognition performance.

### V. CONCLUSIONS

We have presented our initial effort for a fully automated analysis of images acquired by our custom-built high resolution towed digital imaging system (ISIIS). As we are interested in the identification and quantification of a large number of organisms, some of which are relatively similar to each other, our method aims at (i) the detection and (ii) the
recognition of specimens of interest. To test the performance of our integrated system, we have built a library of a limited number of taxa, and we presented our detection and recognition results.

We are currently enriching our library with more samples of the existing taxa, as well as including more taxa of interest. Also, we are working on a more efficient recognition approach that can achieve low computational times. Our ultimate goal is to build a fully automated system that can detect and recognize taxa online, i.e., in situ, while acquiring the images.

<table>
<thead>
<tr>
<th>Region</th>
<th>Specimen</th>
<th>Number of keypoints detected</th>
<th>Detected matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>input</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trichodendrium</td>
<td>13</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Larvae</td>
<td>6</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Fish larvae</td>
<td>34</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Copepod</td>
<td>12</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Copepod</td>
<td>10</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Chaetognaths</td>
<td>31</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Chaetognaths</td>
<td>31</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 5. ROC curves along with the recognition rates for each taxon from our library; x-axis indicates the False Positive Rate (FPR) (or specificity), while the y-axis indicates the True Positive Rate (TPR) (or sensitivity).

### REFERENCES


