

RESEARCH

Open Access



# Deriving image features for autonomous classification from time-series recurrence plots

Jan Schulz\*, Andrea Mentges and Oliver Zielinski

## Summary

This paper shows the use of a specific type of time series analyses, the so named recurrence plot (RP), for investigations of the outer hull of an imaged and pre-segmented object to derive image features suitable for usage in classifiers. Additionally to the features derived by the well documented recurrence quantification analysis (RQA) a new set of features was developed based on closed structures ("eyes") in a RP. The new features were named eye structure quantification (ESQ). Two sets of images are analysed: a) 1023 in-situ plankton images comprising nine different organism classes, and b) each 50 algorithmically created geometric shapes of five different classes. These images were characterised by standard image features, RQA quantification and the newly proposed features. A Linear Discriminant Analysis (LDA) was used to determine discriminative success between the classes of plankton organisms or geometric shapes respectively. The discriminative success was compared between a model using standard features and additional RQA and ESQ. For the high intra- and low interclass variance of the plankton contour line data set the included features enhanced discriminative success by 3 % to a maximum of 65.8 %. For the data set of geometric shapes an increase of 6.8 % to 95.2 % was observed. Although the overall increase of discriminative success was not extraordinary high by using a linear model, it can be seen that both RQA and ESQ are valuable auxiliary features to split specific classes from the entire population. Thus, they may also be valuable for methods mapping the finite dimensional feature space into higher dimensional spaces (e.g. Kernel trick, Support Vector Machines).

## Background

Time series are sequences of metered values. Such readings generally have a natural chronology, are non-circular and exhibit a defined start and end for the recorded time interval. Typical examples of time series are e.g. tidal signals, meteorological observations, stock exchange quotations or cardiograms. Tools for the investigation of time series include a large portfolio of forecasting, estimating or classifying methods and the identification of dependencies, harmonic anomalies or recurrences.

Especially the identification of recurrences allows identifying whether the current state of a dynamic system retraces prior observed states. Eckmann et al. [1] introduced a visual method to investigate such recurrences. The respective tool is the recurrence plot (RP). It uses the time delay embedding theorem (DET, [2]) to display previously encountered states in a phase space. Advantageously RPs

using DET not only identify parity situations but also approximations to the compared template structure with given precision. Thus, a RP identifies sections of phase space trajectories that converge. The recurrence quantification analysis (RQA) comprises a set of heuristically developed methods to derive numerical characterisations of the complexity of a RP and its small-scale features (e.g. [3–5]). Here we first investigate the use of RP and RQA for automated image discrimination and apply it to the very different field of marine plankton data.

For a wide range of marine investigations it is important to chart distribution, abundance and diversity of major plankton groups and suspended material. Traditional methods include sampling the water column by nets of fine gauze and defined mouth opening. Skilled taxonomists determine and enumerate biota from aliquots under stereomicroscopes. The human eye easily gives a first taxonomic impression based on shape and habitus of an organism. Manual arrangement to best see specific morphological

\* Correspondence: jan.schulz@uni-oldenburg.de  
Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, 26111 Oldenburg, Germany

characteristics (e.g. bristles, setae or body appendages) further allows a more precise taxonomic identification. Even if an object cannot be determined to species level a super-ordinate taxonomic group membership can be assigned; often sufficient for the scientific question at hand.

During the last decades various in-situ plankton imaging systems were developed. Today most of these devices are capable to sufficiently image tiny organisms or particles for detailed analyses (e.g. [6]). Although accurate species identification often fails, the major taxonomic group membership can generally be determined. Thus, these approaches add new opportunities to net samplings and have proven to be valuable tools (e.g. [7]). By this, they can partially substitute labour and cost-intensive net analyses and continuously map fine-scale distributions of dispersed objects in the water column.

However, the sheer amount of images and data face researchers with new challenges. In contrast to net samplings in situ systems deliver two dimensional still images which represent information of incident light scattered from imaged objects at arbitrary angles and spatial alignment. Although alignment can be partially controlled by fluidic design of the sampling chambers object appearances are still highly variable (e.g. clinging or abducted antenna and body appendages). To fully utilise the advantages of in-situ plankton imaging systems requires sophisticated machine vision approaches aiding researchers to handle the flood of information. For this, automatic image feature extraction and classification are required that are capable to assign major group memberships in a comparable way as a human taxonomist would.

A variety of algorithms are available to extract numerical features from 2D images and their silhouettes. Standard methods are moments derived from pattern intensity variations, colour information and geometric parameters, like roundness, compactness or elliptical shape equivalents. More sophisticated methods investigate contour lines by Fourier descriptors (e.g. [8]), characteristic inflexions (e.g. [9]) or identification of points of interest in scale space (e.g. [10, 11]).

Although, such features are generally invariant to scale, rotation and translation downstream classification systems often lack high discriminatory power for plankton specimens (e.g. [12]). An important factor is the multivariate high intra-class heteroscedasticity. This high variability is a general challenge when compiling feature sets considering contour lines of plankton species. Depending on illumination, resolution, contrast and orientation the outer contour and tissues appear highly variable. This arises from on-site illumination variations and flexibility and agility of body parts and appendages. Thus, predictability of the contour line's curve progression is comparable with dynamic systems.

Here we present an approach to apply the recurrence plot method on circular contour line data by using a modified embedding, where the contour line data are augmented by

recycled elements. The resulting RP is the basis to get a first glimpse about usefulness of RQA scalars as features for automated classification systems. For comparison we used two different image sets. The first set is composed of geometric forms, while the second is compiled from images of plankton specimens and marine snow taken under arbitrary angles and showing high morphological variability.

## Methods

### Images

#### *Geometric shapes*

Two sets of images were used. The first is a generic set of algorithmically created geometric shapes. This data set includes 50 shapes each out of five classes: circles, ellipses, squares, rectangles and triangles (Appendix A: Fig. 4). To minimise the impact of the contour line length the shapes were chosen to have a comparable intra-group perimeter (mean 140.57, SD 1.13).

#### *Plankton images*

The second image set contains 1023 images out of 21 groups (Table 1). These 21 groups can finally be super-ordinated into 9 higher classes. They present mainly taxonomic or morphological plankton groups and marine snow. This data set is published and freely accessible via the Pangaea data publisher system [13].

Images were sampled with the Lightframe On-sight Key-species Investigation (LOKI) system [6]. The advantage of the LOKI sampling design is the high contrast imaging of minute objects at high magnifications (here  $\sim 15 \mu\text{m}$  per pixel) at very short shutter times ( $< 30 \mu\text{s}$ ) in a physically constrained volume, being transparent before and behind the depth of field. Thus, the system delivers bright and detailed images of taxons that are often destroyed during traditional net samplings. Images were manually classified by declared experts of the respective plankton taxon. The images were taken from a larger subset sampled during an earlier expedition off the coast of Peru (rf. [14]) and represent major plankton classes of the on-site community.

#### **Standard image features**

The 8 image features, hereafter referred to as *STANDARD* (Table 2), were extracted by using the MATLAB function '*regionprops*' and '*graycoprops*'. For more detailed information see MATLAB documentation. **Area:** Number of pixels within the object's contour line. **Compactness:** Quantified by the inverse Patton Shape Index [15], which compares the perimeter of the shape to the perimeter of a standard shape. An index of 1 equals a perfect circle. **Contrast:** Intensity contrast between neighbouring pixels (zero for constant images). **Eccentricity:** Eccentricity of an ellipse corresponding best to the object shape. **Hu Moments:** The seven moment invariants of the object [16], calculated using a script by Gonzalez et al. [17]. In the following the

**Table 1** Taxonomic class sizes used in the analyses

Taxon/Class	#	Total #
Annelida		50
- Polychaeta	50	
Appendicularia		50
- Oikopleura	50	
Bacillariophyceae		100
- Coscinodiscales	50	
- Rhizosoleniales	50	
Cnidaria		60
- Medusae	30	
- Siphonophora	30	
Crustacea		515
- Amphipoda	50	50
- Copepoda		
o Calanidae		125
▪ Acartia	50	
▪ Calanus	50	
▪ Calocalanus	25	
o Cyclopoidae		50
▪ Oithona	50	
o Poecilostomatoida		140
▪ Corycaeus	50	
▪ Oncaea	50	
▪ Sapphyrina	40	
- Euphausiacea		50
o Genera not further separated	50	
- Ostracoda		50
o Genera not further separated	50	
- Nauplii		50
o Various genera and species	50	
Dinoflagellata		45
- Noctiluca	45	
Marine snow		150
- Heterogeneous marine snow particles	150	
Mollusca		28
- Gastropoda	28	
Vertebrata		25
- Fish larvae	25	
Total		1023

first Hu-moment was used only, as higher moments sometimes caused collinearities in the following analyses. **Homogeneity**: Closeness of the distribution of elements in the normalised grey-level co-occurrence matrix to its diagonal (one for diagonal matrix). **Perimeter**: The perimeter of the organism's shape in pixels. **Solidity**: Quotient of the

**Table 2** Categories of numerical features extracted for each image

STANDARD	RECURRENCE QUANTIFICATION ANALYSIS (RQA)	EYE STRUCTURE QUANTIFICATION (ESQ)
Area	Clustering coefficient	Mean eye size
Compactness	Determinism	Median pixels of eye
Contrast	Entropy diagonal length	Number of eyes
Eccentricity	Laminarity	Summed pixel in eyes
Hu1-Moment	Longest diagonal length	
Homogeneity	Longest vertical length	
LengthBoundary	Mean diagonal length	
Solidity	Recurrence period density	
	Recurrence rate	
	Recurrence time1	
	Recurrence time 2	
	Transitivity	
	Trapping time	

All features of a category have been either used in the analyses or excluded

number of pixels within the object contour line and the number of pixels in the respective convex hull.

#### Contour line extraction and measurement

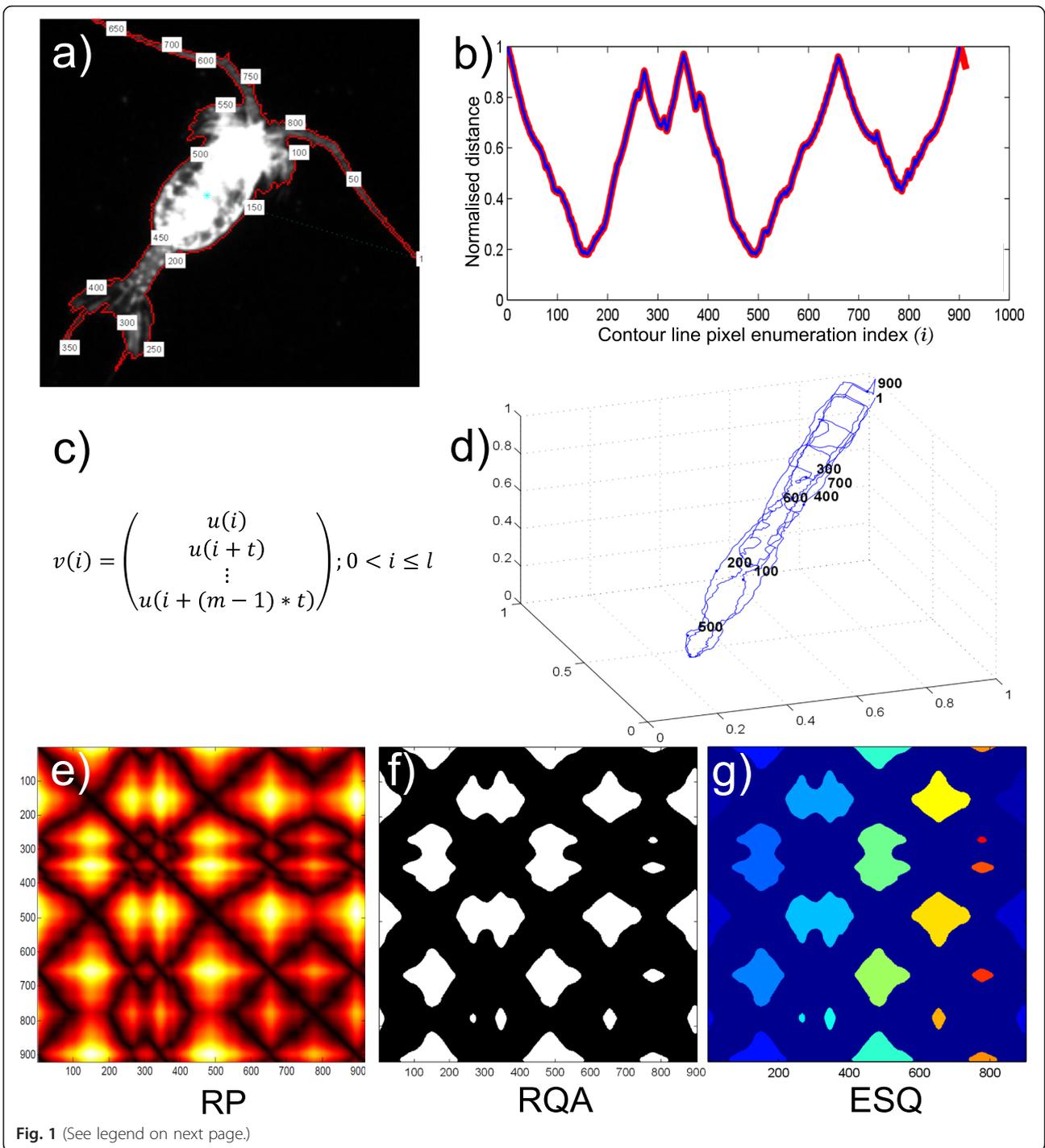
For each imaged object the coordinates of the mass centroid is calculated. Additionally, the finite contour outline of the organism is determined. The contour line is a list of length  $l$  giving the coordinates of the points at the organism's outer boundary (Fig. 1a). From the centroid the distance to each point with index  $i$  of the contour line is calculated clockwise according to a pre-defined norm. In the following the Euclidean norm was used. Values are tabulated in a list  $u$  (Eq. 1) and normalised to 1:

$$u(i); 0 < i \leq l \quad (1)$$

The basis for the recurrence quantification analyses thus is a list of distances  $u$ , from each contour line point to the centroid. The list is shifted in a way that the first index  $u$  (1) represents the maximum distance found; increasing indices clockwise enumerate the subsequent distances (Fig. 1b).

#### Embedding

Using the embedding theorem [2] a phase space trajectory in dimension  $m$  with  $m > 1$  is created from  $u$ . Therefore  $m$  values from  $u$  are used to create a new vector  $v$  of dimension  $m$  representing the points of the phase space trajectory. Values used from  $u$  are chosen to have equidistant spacing  $t$ . As mentioned before the contour line data, in contrast to a time series, represents a circular structure. Therefore the first  $(m-1)*t$  elements of  $u$  need to be recycled and added to the end of the list  $u$ . The length of  $u$  becomes  $l + (m-1)*t$ . In case of  $(m-1)*t > l$  the elements of  $u$  need to be re-



(See figure on previous page.)

**Fig. 1** Schematic workflow. **a**) Extraction of the outer contour line of the object (*red line*). The cyan dot indicates the mass centroid. **b**) For each point of the red contour line the distance to the centroid is measured according to a predefined norm and normalised. The greatest distance is stored as first element in a list  $u(1)$ . All other distances  $u(i)$  are enumerated clockwise from this starting point (*blue line*). The red line is the distance list augmented by  $(m-1)*t$  elements, recycling the beginning of  $u$ . Parameters  $m$  and  $t$  are given by the subsequent embedding. **c**) From list  $u(i)$  a set of  $m$  dimensional vectors is derived, each having  $m$  elements of  $u$  with an equidistant spacing of  $t$ . The chronology of  $u(i)$  is embedded in  $v(i)$ . **d**) A phase space trajectory in  $m$  dimensional space can be constructed from  $v$  (here shown for an example with  $m=3$ ). Numbers attached to some points of the phase space trajectory refer to index  $i$  of the original contour line. **e**) For each point  $i$  of the phase space trajectory the distance to any other point  $j$  is measured and tabulated. This can be plotted as a colour heat map. **f**) In a later step it is checked, whether the respective distance is greater than a given threshold  $\varepsilon$  (Heaviside operator). The result is tabulated as a square, symmetric and binary matrix, the recurrence plot (RP). White dots indicate that the distance between  $v(i)$  and  $v(j)$  is greater than  $\varepsilon$ . On the main diagonal points are compared against themselves. Thus, the distance is always zero. From the RP a number of numerical features are derived in the subsequent recurrence quantification analysis (RQA, refer to the text). **g**) The enclosed white coherent areas within a RP have been termed “eyes”. Due to the circular data structure and above mentioned augmentation the truncated eyes along the borders need to be interpreted as connected structures on the opposite sides of the plot. This is displayed by matching colours of associated eyes. This plot serves as a basis for the eye structure quantification (ESQ, see text)

recycled. This results in a set of vectors  $v$  defining the points of the phase space trajectory (Equation 2):

$$v(i) = \begin{pmatrix} u(i) \\ u(i+t) \\ u(i+(m-1)*t) \end{pmatrix}; 0 < i \leq l \quad (2)$$

Dimension  $m$  and time delay  $t$  have to be chosen properly prior to analysis. To investigate their impact several tests were performed beforehand for  $1 < m \leq 10$  and  $1 \leq t \leq 10$ . For the examples given in this paper  $m=6$ ;  $t=6$ ;  $\varepsilon=3.0$  was used. Sample plots for various parameter combinations are given in Appendix B: Figs. 5, 6 and 7.

### RP - Recurrence plot

For each investigated object a matrix  $R$  is calculated from the phase space trajectory (Fig. 1d). For each element  $R(i,j)$  the norm  $\| \cdot \|$  between the vectors  $v(i)$  and  $v(j)$  is calculated (Eq. 3). For the results presented in this paper the Euclidean norm was chosen. Finally,  $R$  is a  $l \times l$  square and symmetric matrix that can be displayed as a false colour heat map representing the distances between all points of the phase space trajectory according to the used norm (Fig. 1). For downstream processing the Heaviside step function  $\Theta(\cdot)$  is applied to identify those distances of phase space trajectory points that fall below a predefined minimum value  $\varepsilon$ . Thus, the definition of the recurrence plot becomes a matrix of binary values given by:

$$R(i,j) = \Theta(\varepsilon - \|v(i) - v(j)\|); 1 \leq i \leq l \text{ and } 1 \leq j \leq l \quad (3)$$

Consequently, the main diagonal of such a recurrence plot represents the distance of a point to itself and is therefore 0. Once the Heaviside step function was applied all off-diagonal non-zero entries of  $R$  indicate phase space approximations smaller than  $\varepsilon$  having a distance on the contour line of  $li-jl$ .

Side diagonals parallel to the main diagonal indicate that structures of the contour line are similar in phase space. The length of the similarity structure is equivalent to the length along the axis, with the latter given distance on the

contour line. Among diagonal structures coherent areas exceeding  $\varepsilon$  (name “eyes”) can be found (Fig. 1e-f). These patterns within a RP represent major characteristics and are investigated in detail numerically.

### RQA - Recurrence quantification analysis

Parameters of the Recurrence Quantification Analysis (RQA, Table 2) were obtained using the Cross Recurrence Plot Toolbox [5, 18]. Values transferred in the function call are the embedding vectors  $v(i)$ , dimension  $m$ , time delay  $t$ , size of neighbourhood  $\varepsilon$  and norm to be used (Euclidean). A total of 13 features were extracted from each RP (Table 2). Details are given in [3–5, 19] or [20]: **Clustering coefficient** gives the degree to which points of the phase space trajectory tend to cluster. **Determinism** gives the proportion of recurrent points forming diagonals. **Entropy diagonal length** gives the Shannon entropy of the probability distribution of the lengths of the diagonals. **Laminarity** gives the amount of recurrence points forming vertical structures. **Longest diagonal length** gives the counted length of the longest diagonal. **Longest vertical length** gives the counted length of the longest vertical. **Mean diagonal length** gives the average length of the diagonal structures. **Recurrence period density** gives the periodicity of the signal in the RP. **Recurrence rate** gives the density of observed recurrence points in the RP. **Recurrence times** give an estimation of the periodicity in the RP signal. **Transitivity** gives the probability that two points of the phase space trajectory neighbouring a third are also directly connected. **Trapping time** gives the average length of the vertical structures.

### ESQ - Eye structure quantification

From the recurrence plot matrix  $R$  additional features were derived. In the *Eye Structure Quantification (ESQ)* distribution and size of enclosed structures, the so-called ‘eyes’, were measured. Due to the circular structure of an organism’s contour line opposite sides of the RP need to be interpreted as connected structures. Thus, eyes truncated at the borders of  $R$  have to be associated with their counterpart on

the opposite side prior to evaluation (Fig. 1g). After identification of associated eyes, the total number of eyes, mean number of pixels per eye (e.g. mean eye size), the median of the numbers of pixels per eye, and total number of pixels in all eyes were determined. Increasing eye numbers generally indicate, that a high number of independent features recurrences in phase space are found. These are often associated with repetitive morphological structures of the object, like polychaete parapodia, silica spicules or regular diatom frustule indentations.

#### LDA - Linear discriminant analysis

A Linear Discriminant Analysis (LDA, [21–23]) was used as classifier. The LDA model was built with the training data set (geometric shapes or plankton images) and tested against itself to investigate the role of the included features. An individual LDA was run for each of the 4 feature combinations (Table 3) and both image sets. LDA results evaluated in this paper are:

- Coefficients of linear discriminant roots. These values represent the loadings and thus, importance of the individual features during discrimination.
- Proportions of trace. These values give the variance explained by the respective root. As explained variance decreases with each successive root we give just the first roots in this paper; although for some LDA's more roots could be given (number of roots equals number of objects or number of included features minus one; whatever is lower).
- Confusion matrices. They show the rate of true positive and false positive classifications.
- From the coefficients of linear discriminants, the most important features were identified that best separate objects by the respective root. A feature was considered to be important when it's loading reached at least 10 % of the maximum feature loading on either side of a root's spanned hyperplane.
- Canonical scores. The scores of the individual objects were plotted to visualise the discriminative success among object classes for the respective roots.

#### Computational work

Image processing and feature extraction (RQA, ESQ) were performed in Matlab (MathWorks, 2013, v8.1.0.604). The

**Table 3** Feature combinations used for the different LDA models

LDA setup	Included features
1	<i>STANDARD</i>
2	<i>RQA</i>
3	<i>STANDARD &amp; RQA</i>
4	<i>STANDARD &amp; RQA &amp; ESQ</i>

Each of the four models was run individually for the shapes and the plankton image data set

LDA models were implemented in R ([www.r-project.org](http://www.r-project.org)), using the additional package MASS.

## Results

### LDA - Linear Discriminant Analysis

#### Geometric shapes

**Standard** First LDA included the *STANDARD* parameters. The first discriminant root (LD1) explained 72.58 % of the observed variance, while the second (LD2) explained additional 26.02 % (Table 4). LD3 and LD4 are of less importance, as their cumulative impact is less than 1.5 %. It is obvious, that parameters like *Area* and *Contrast* have least impact for discriminating geometric structures. The confusion matrix shows that 88.4 % of the geometric shapes were classified correctly (Table 5). In the canonical plot (Fig. 2a) rectangles and circles show a clear clustering tendency, while other geometric shape categories show much higher dispersal.

**RQA** The second LDA included the *RQA* toolbox parameters, where LD1 explains 73.42 % of the observed variance and LD2 explains 14.13 % (Table 6). The cumulative explanatory power of LD3 and LD4 still comprises approximately 12.5 %. The confusion matrix shows a total discrimination success of 83.6 % (Table 5). It can be seen in the canonical plot, that rectangles and circles separate from other categories (Fig. 2b) but inter-class discrimination is lower compared to *STANDARD*. The three other classes separate well, but show a higher dispersal on both roots.

**Standard & RQA** The third LDA included both the *STANDARD* and *RQA* parameters. LD1 explains 59.82 % of the observed variance, while LD2 contributes with a value as high as 31.63 % (Table 7). Again LD3 and LD4 have neglectable explanatory power. The confusion matrix of the model shows a discrimination success of 95.6 % (Table 5). In the canonical plot the classes show a well discriminable clustering (Fig. 2c).

**Standard, RQA & ESQ** The fourth LDA included the *STANDARD*, the *RQA* and the newly developed parameters *ESQ*. Again the first two roots show highest proportions of trace (Table 8), with LD1 explaining 62.32 % of the observed variance and LD2 explaining 29.84 %. The confusion matrix shows a discrimination success of 95.2 %. In the canonical plot the classes again show a well discriminable clustering (Fig. 2d). Although some minor differences are observable, the result is comparable to the latter *STANDARD & RQA* setup.

### Plankton images

#### Standard

The first discriminant root (LD1) explained 51.93 % of the observed variance, while the second (LD2) contributed

**Table 4** LDA Geometric shapes

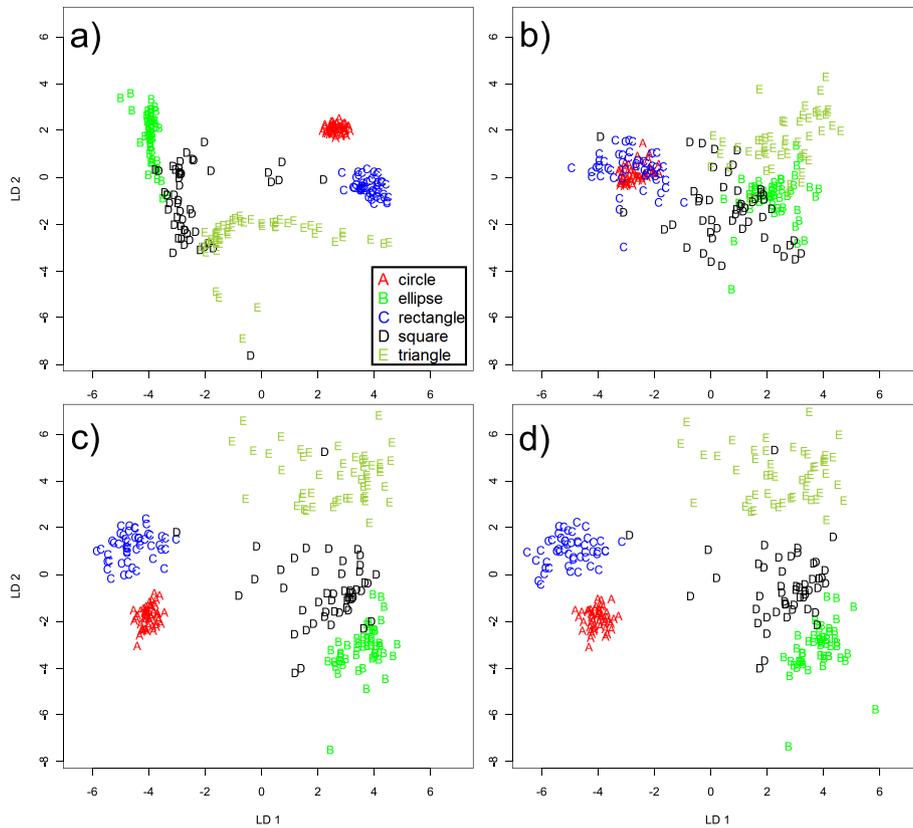
<i>STANDARD</i>	LD1	LD2	LD3	LD4
<i>Area</i>	-6.9626 e-05	2.3917 e-05	1.9991e-04	3.4263e-04
<i>Compact</i>	-1.1407 e+01	2.8052 e+01	-1.0988 e+01	-1.2826 e+01
<i>Contrast</i>	-1.4711 e-06	-4.2640 e-07	8.6186 e-06	-3.2337 e-05
<i>Eccentricity</i>	-9.6114 e+00	2.4681 e+00	5.3053 e-01	-1.1433 e+00
<i>HU1</i>	6.7315 e+00	-1.1663 e+01	-6.4077 e+01	1.4417 e+01
<i>Homogeneity</i>	3.4927 e+00	2.7952 e+01	-7.0628 e+01	1.7027 e+00
<i>LengthBoundary</i>	4.2602 e-03	2.8937 e-04	-2.5557 e-02	-1.6500 e-02
<i>Solidity</i>	9.4545 e+00	-3.9641 e+01	1.4252 e+01	3.5592 e+01
Proportion of Trace	0.7258	0.2602	0.0105	0.0035

Loading coefficients of the linear discriminants using the geometric image set and the *STANDARD* parameters

**Table 5** LDA Geometric shapes

<i>STANDARD</i>	Circle	Ellipse	Rectangle	Square	Triangle	Prediction success
Circle	<b>50</b>	0	0	0	0	1.00
Ellipse	0	<b>45</b>	0	5	0	0.90
Rectangle	0	0	<b>50</b>	0	0	1.00
Square	1	6	1	<b>37</b>	5	0.74
Triangle	0	0	7	4	<b>39</b>	0.78
Model						<b>0.884</b>
<b>RQA</b>						
Circle	<b>47</b>	0	3	0	0	0.94
Ellipse	0	<b>41</b>	0	6	3	0.82
Rectangle	1	0	<b>48</b>	1	0	0.96
Square	1	10	6	<b>31</b>	2	0.62
Triangle	1	5	0	2	<b>42</b>	0.84
Model						<b>0.836</b>
<b>STANDARD &amp; RQA</b>						
Circle	<b>50</b>	0	0	0	0	1.00
Ellipse	0	<b>46</b>	0	4	0	0.94
Rectangle	0	0	<b>50</b>	0	0	1.00
Square	1	1	2	<b>43</b>	3	0.88
Triangle	0	0	0	0	<b>50</b>	1.00
Model						<b>0.956</b>
<b>STANDARD &amp; RQA &amp; ESQ</b>						
Circle	<b>50</b>	0	0	0	0	1.00
Ellipse	0	<b>46</b>	0	4	0	0.92
Rectangle	0	4	<b>49</b>	1	0	0.90
Square	1	3	1	<b>43</b>	2	0.86
Triangle	0	0	0	0	<b>50</b>	1.00
Model						<b>0.952</b>

Confusion matrices of the different parameter combinations used for the LDA analyses of geometric shapes  
 Values in bold represent true positive classifications of the LDA model and overall discriminative success



**Fig. 2** LDA Geometric shapes. Canonical plot of the linear discriminants. The parameters used for the analyses were **a)** STANDARD, **b)** RQA, **c)** STANDARD & RQA and **d)** STANDARD & RQA & ESQ

**Table 6** LDA Geometric shapes

RQA	LD1	LD2	LD3	LD4
Clustering coefficient	-3.3203 e-02	-7.3117 e+01	-2.8931 e+01	-6.0119 e+01
Determinism	1.4724 e+02	-4.3258 e+01	2.0517 e+01	6.1069 e+01
Entropy Diagonal Length	1.5459 e+00	7.7545 e-01	9.2093 e-01	-1.4666 e+00
Laminarity	-2.8086 e+02	1.0565 e+01	1.1170 e+00	-2.3284 e+02
Longest Diagonal Length	-2.2921 e-02	1.2801 e-02	-1.7178 e-02	-5.2463 e-03
Longest Vertical Length	-9.7082 e-03	-1.7140 e-03	5.9098 e-03	2.1794 e-03
Mean Diagonal Length	-2.2354 e-01	2.4054 e-02	-2.8575 e-02	-3.5738 e-01
Recurrence Period Density	-1.6803 e+00	5.0734 e+00	4.5539 e+00	7.5552 e+00
Recurrence Rate	1.2637 e+01	-6.0927 e+00	1.3810 e+01	-6.1760 e+01
Recurrence Time1	-1.8844 e+01	6.3186 e+01	3.1819 e+01	-1.8392 e+01
Recurrence Time2	6.3736 e-02	-1.7368 e-01	5.0725 e-02	-1.7492 e-01
Transitivity	-3.3545 e+01	1.0021 e+02	7.0535 e+01	5.9794 e+01
Trapping Time	2.6801 e-01	1.3032 e-01	-7.1610 e-03	6.1896 e-01
Proportion of Trace	0.7342	0.1413	0.0918	0.0327

Loading coefficients of the linear discriminants using the geometric image set and the RQA toolbox parameters

**Table 7** LDA Geometric shapes

STANDARD & RQA	LD1	LD2	LD3	LD4
Area	-5,9401E-05	-1,9400E-04	-6,2879E-05	1,0076E-04
Compact	3,5423E + 00	-2,5172E + 01	1,1080E + 01	9,1550E-01
Contrast	2,9356E-06	-4,7963E-06	-2,9805E-06	-6,5782E-06
Eccentricity	7,2951E + 00	-6,3447E + 00	-2,1702E + 00	-9,3084E-01
HU1	1,5031E + 01	5,4044E + 01	1,4604E + 01	-1,5518E + 01
Homogeneity	-1,4477E + 01	-2,6754E + 01	1,4672E-01	-3,4403E + 00
LengthBoundary	-2,5076E-03	1,1792E-02	-1,6241E-03	-2,6010E-03
Solidity	1,4425E + 00	3,5392E + 01	-1,7666E + 01	3,9915E + 00
Clustering coefficient	-2,6156E + 01	-2,8633E + 01	-3,7726E + 01	-7,8372E + 00
Determinism	1,5009E + 01	-7,2239E + 01	8,2244E + 01	5,3571E + 01
Entropy diagonal length	1,1423E + 00	2,3649E + 00	1,8728E + 00	-1,3749E + 00
Laminarity	-5,2600E + 01	7,8644E + 01	-1,4305E + 02	-2,0549E + 02
Longest diagonal length	-2,5076E-03	1,1792E-02	-1,6241E-03	-2,6010E-03
Longest vertical length	-5,8332E-03	-6,5336E-03	-4,0279E-03	-4,7670E-03
Mean diagonal length	-2,0581E-01	-9,1709E-02	-1,3184E-01	-1,6458E-01
Period density	5,9440E-01	-4,4322E + 00	3,2363E + 00	2,4944E + 00
Recurrence rate	3,3347E + 01	1,4299E + 01	-1,5627E + 01	-4,9725E + 01
RecurrenceTime1	4,3389E + 00	2,6093E + 00	-1,1137E + 00	-4,4655E + 01
RecurrenceTime2	1,2758E-02	-1,2454E-01	-7,1585E-02	-1,0382E-01
Transitivity	1,9919E + 01	4,1271E + 01	5,6892E + 01	-3,2801E + 01
TrappingTime	2,4661E-01	2,3180E-01	2,6364E-01	3,2499E-01
Proportion of Trace	0.5982	0.3163	0.0532	0.0323

Loading coefficients of the linear discriminants using the geometric image set and the *STANDARD* and *RQA* parameters

with 28.97 % (Table 9). Cumulated proportions of trace of LD3 and LD4 explain less than 16 %. The confusion matrix (Table 10) shows a total discrimination success of 62.8 %. The canonical plot (Fig. 3a) reveals good separation between some classes. Dinoflagellata and Bacillariophyceae separate well from Appendicularia, Vertebrates and Cnidarians. The majority of Crustacea, Annelida, and Mollusca overlap largely with Marine snow.

#### RQA

The first discriminant root (LD1) explained 60.66 % of the observed variance, while the second (LD2) contributed with an additional 19.77 % (Table 11). Cumulated LD3 and LD4 contributed with less than 16 %. The confusion matrix (Table 10) shows a total discrimination success of 55.0 %. As in the previous LDA, centroids of Bacillariophyceae and Dinoflagellata separate from the majority of objects in the canonical plot (Fig. 3b). The same is observed for Appendicularia and Vertebrata, although separation on LD2 is more pronounced, than in the previous plot on LD1.

#### Standard, RQA

The first discriminant root (LD1) explained 41.58 % of the observed variance, while the second root (LD2)

contributed with 31.35 % (Table 12). The cumulative explanatory power of LD3 and LD4 was 20.78 %. The confusion matrix shows a total discriminative success of 66.1 %. In the canonical plot (Fig. 3c) it can be seen that centroids of the formerly identified classes (Bacillariophyceae, Dinoflagellata, Appendicularia and Vertebrata) again separate but are now more spread out in the LD1/LD2 plane, allowing better discrimination.

#### Standard, RQA, ESQ

The first discriminant root (LD1) explained 40.15 % of the observed variance, while the second (LD2) explained 32.18 % (Table 13). Roots LD3 and LD4 contributed with a cumulative observed variance of 20.88 %. In the confusion matrix the total discriminative success is found to be 65.8 % (Table 10). The canonical plot (Fig. 3d) is comparable to the previous LDA (*STANDARD & RQA*), but shows a slight shift of Vertebrata, better separating from the remaining classes.

#### Importance of the image features

##### Geometric shapes

The features with highest loadings for LDA image classification of the geometric shapes are listed in Table 14. The most frequently occurring features are *Transitivity*, *Determinism*,

**Table 8** LDA Geometric shapes

<i>STANDARD, RQA &amp; ESQ</i>	LD1	LD2	LD3	LD4
<i>Area</i>	-1.1784 e-04	-1.9232 e-04	-1.1059 e-04	1.9359 e-04
<i>Compact</i>	3.4301 e+00	-2.5118 e+01	1.1435 e+01	2.4919 e+00
<i>Contrast</i>	1.5530 e-06	-3.7785 e-06	-5.2896 e-06	-3.2900 e-06
<i>Eccentricity</i>	7.7392 e+00	-6.2394 e+00	-2.3488 e+00	-5.5377 e-01
<i>HU1</i>	-1.1736 e+01	5.6397 e+01	1.5916 e+01	1.0685 e+01
<i>Homogeneity</i>	-1.7666 e+01	-2.6344 e+01	1.4941 e+00	-2.1289 e+00
<i>Length Boundary</i>	-1.7489 e-02	1.2571 e-02	-5.3791 e-03	1.4207 e-02
<i>Solidity</i>	-1.4304 e+00	3.5787 e+01	-1.9367 e+01	4.7933 e+00
<i>Clustering Coefficient</i>	-3.6658 e+01	-2.5735 e+01	-3.8500 e+01	-7.0815 e+00
<i>Determinism</i>	2.5800 e+01	-8.5462 e+01	9.3743 e+01	5.6418 e+01
<i>Entropy diagonal Length</i>	1.2321 e+00	2.4068 e+00	2.1062 e+00	-1.3100 e+00
<i>Laminarity</i>	-3.7915 e+01	1.0449 e+02	-1.6604 e+02	-2.3688 e+02
<i>Longest diagonal Length</i>	-1.7489 e-02	1.2571 e-02	-5.3791 e-03	1.4207 e-02
<i>Longest vertical Length</i>	-6.1717 e-03	-7.0950 e-03	-4.9047 e-03	-3.3528 e-03
<i>Mean diagonal Length</i>	-2.0822 e-01	-7.9261 e-02	-9.9609 e-02	-2.1448 e-01
<i>Recurrence period density</i>	7.2700 e-01	-5.8102 e+00	3.4804 e+00	4.6755 e+00
<i>Recurrence rate</i>	4.2988 e+01	1.4841 e+01	-1.8897 e+01	-4.9004 e+01
<i>Recurrence time1</i>	7.8288 e+00	3.4506 e+00	2.8467 e+00	-4.5792 e+01
<i>Recurrence time2</i>	1.3043 e-02	-1.1242 e-01	-4.3600 e-02	-1.3724 e-01
<i>Transitivity</i>	3.0299 e+01	4.0309 e+01	6.5165 e+01	-3.5768 e+01
<i>Trapping time</i>	2.7737 e-01	2.1055 e-01	2.2561 e-01	3.5996 e-01
<i>Mean size eyes</i>	-5.0701 e-04	-5.3573 e-05	-1.5723 e-04	3.5322 e-04
<i>Median pixels in eyes</i>	3.2251 e-04	-4.0013 e-05	-4.4404 e-05	1.2952 e-04
<i>Num eyes</i>	2.0862 e-03	-3.9122 e-03	8.4803 e-03	-1.8939 e-03
<i>Sum pixels in eyes</i>	1.3377 e-04	-1.0762 e-06	4.1416 e-05	-1.6796 e-04
Proportion of Trace	0.6232	0.2984	0.0490	0.0294

Loading coefficients of the linear discriminants using the geometric image set and the *STANDARD, RQA* and *ESQ* parameters**Table 9** LDA Plankton images

<i>STANDARD</i>	LD1	LD2	LD3	LD4
<i>Area</i>	9,5536E-06	1,2253E-05	-2,0796E-05	2,2256E-05
<i>Compact</i>	3,5030E+00	-1,7440E+00	7,3339E+00	8,8077E+00
<i>Contrast</i>	-6,3512E-06	-2,5975E-06	1,8473E-05	-1,3421E-05
<i>Eccentricity</i>	-4,2015E+00	1,7316E+00	1,6618E+00	-1,1651E+00
<i>HU1</i>	8,9741E+02	2,9666E+02	4,1634E+01	-2,2460E+02
<i>Homogeneity</i>	-9,5771E+00	-2,1946E+01	-6,0204E+00	-1,8834E+01
<i>Length boundary</i>	-5,3401E-04	-5,7619E-04	-8,0676E-04	7,3831E-04
<i>Solidity</i>	1,6513E+00	-8,8657E-01	-8,2544E+00	-8,1160E+00
Proportion of Trace	0.5193	0.2897	0.1026	0.0534

Loading coefficients of the linear discriminants using the *STANDARD* features

**Table 10** LDA Plankton images

<b>STANDARD</b>	Annelida	Appendicularia	Bacillariophyceae	Cnidaria	Crustacea	Dinoflagellata	Marine snow	Mollusca	Vertebrata	Prediction success
Annelida	<b>15</b>	1	0	2	30	0	0	0	2	0.30
Appendicularia	0	<b>37</b>	0	0	8	1	0	0	4	0.74
Bacillariophyceae	0	0	<b>34</b>	0	8	16	37	5	0	0.34
Cnidaria	8	9	0	<b>29</b>	8	2	3	1	0	0.48
Crustacea	21	2	6	4	<b>455</b>	0	15	4	3	0.89
Dinoflagellata	0	1	3	0	0	<b>38</b>	0	3	0	0.84
Marine snow	4	2	23	0	103	0	<b>18</b>	0	0	0.12
Mollusca	0	0	8	0	6	3	2	<b>9</b>	0	0.32
Vertebrata	3	10	0	0	8	0	0	0	<b>4</b>	0.16
Model										<b>0.628</b>
<b>RQA</b>										
Annelida	<b>5</b>	1	0	15	29	0	0	0	0	0.10
Appendicularia	0	<b>24</b>	0	0	23	0	0	0	3	0.48
Bacillariophyceae	0	0	<b>35</b>	0	22	26	12	5	0	0.35
Cnidaria	6	2	0	<b>13</b>	33	0	0	0	6	0.22
Crustacea	7	6	10	16	<b>432</b>	0	28	0	11	0.85
Dinoflagellata	0	0	12	0	6	<b>22</b>	4	1	0	0.49
Marine snow	0	0	14	3	107	8	<b>17</b>	1	0	0.11
Mollusca	0	0	4	0	19	2	3	<b>0</b>	0	0.00
Vertebrata	0	8	0	1	4	0	0	0	<b>12</b>	0.48
Model										<b>0.550</b>
<b>STANDARD &amp; RQA</b>										
Annelida	<b>22</b>	1	0	3	24	0	0	0	0	0.44
Appendicularia	0	<b>39</b>	0	0	6	0	0	0	5	0.78
Bacillariophyceae	1	0	<b>48</b>	0	3	20	20	8	0	0.48
Cnidaria	10	1	1	<b>39</b>	5	0	1	2	1	0.65
Crustacea	18	3	6	2	<b>437</b>	0	33	7	4	0.86
Dinoflagellata	0	1	3	0	0	<b>34</b>	0	7	0	0.76
Marine snow	2	2	20	1	86	1	<b>33</b>	5	0	0.22
Mollusca	0	0	5	0	6	2	4	<b>11</b>	0	0.39
Vertebrata	1	7	0	0	7	0	0	0	<b>10</b>	0.40
Model										<b>0.661</b>
<b>STANDARD &amp; RQA &amp; ESQ</b>										
Annelida	<b>24</b>	1	0	3	22	0	0	0	0	0.48
Appendicularia	0	<b>40</b>	0	0	5	0	0	0	5	0.80
Bacillariophyceae	1	0	<b>47</b>	0	2	21	22	7	0	0.47
Cnidaria	6	1	0	<b>36</b>	6	0	4	4	3	0.60
Crustacea	19	4	5	2	<b>436</b>	0	30	7	7	0.85
Dinoflagellata	0	1	3	0	0	<b>33</b>	0	8	0	0.73
Marine snow	2	2	19	1	89	1	<b>31</b>	5	0	0.20

**Table 10** LDA Plankton images (Continued)

Mollusca	0	0	4	0	5	2	5	<b>12</b>	0	0.43
Vertebrata	1	7	0	0	6	0	0	0	<b>11</b>	0.44
Model										<b>0.658</b>

Confusion matrices of the different parameter combinations used for the LDA analyses  
 Values in bold represent true positive classifications of the LDA model and overall discriminative success

*Hu1, Laminarity, Recurrence rate.* Less frequent are *Compact, Clustering coefficient, Solidity,* while *Eccentricity, Homogeneity, Recurrence time 1* and *Recurrence period density* rarely contribute with high loadings to the LDA.

**Plankton Images**

As observed for the set of geometric shapes, a few key features can be identified, which contribute frequently with high loadings to the LDA (Table 15). For the plankton images these are *Laminarity, Hu1* and *Determinism,* followed by *Homogeneity* and *Transitivity.* *Compactness, Clustering coefficient, Recurrence rate, Eccentricity, Entropy diagonal length* and *Recurrence period density* were observed seldomly.

**Discussion**

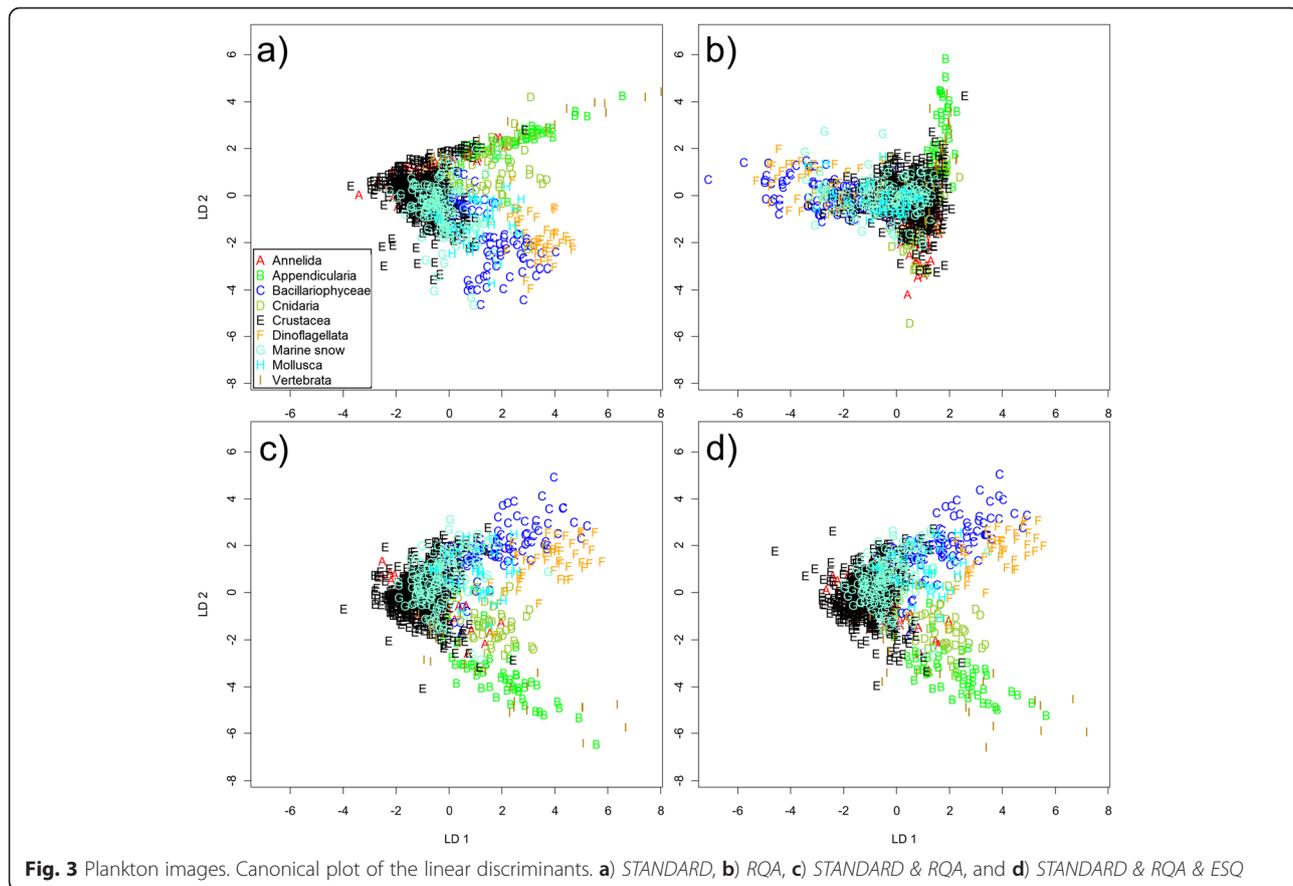
**Method**

The first principle task of this study was to apply the well-established methods of Recurrence Plots (RP) and Recurrence

Quantification Analysis (RQA) in the new context of circular contour line data of an imaged object’s outer hull. To set up the circular contour line data for the proposed methods, each point of the contour line was enumerated and its distance to an arbitrary point was calculated. This arbitrary reference point was static. In contrast to traditional RP and RQA investigations, we augmented the contour line distance data during the embedding process. Thus, the distance data are recycled to allow creating a number of embedding vectors equal to the number of contour lines points. By this, opposite sides of the RP wrap up. This allowed the introduction of the eye structure quantification (ESQ).

**Image discrimination**

The second principle task was to perform an initial test of these methods on both real life plankton data of high contour line variability and a synthetic sample data with similar intra-class structure and symmetry.



**Fig. 3** Plankton images. Canonical plot of the linear discriminants. a) STANDARD, b) RQA, c) STANDARD & RQA, and d) STANDARD & RQA & ESQ

**Table 11** LDA Plankton images

RQA	LD1	LD2	LD3	LD4
Clustering coefficient	9,1646E + 00	1,9776E + 01	-3,6456E + 01	-1,1311E + 01
Determinism	1,2212E + 02	1,3704E + 01	9,8008E + 01	8,7217E + 01
Entropy diagonal length	4,7223E-01	-2,2824E + 00	3,7802E-02	1,0338E + 00
Laminarity	-1,6675E + 02	3,8020E + 01	-1,5864E + 02	-2,2712E + 02
Longest diagonal length	-2,2621E-04	-5,7530E-04	-8,2800E-04	3,0824E-04
Longest vertical length	-4,4431E-04	2,7945E-03	4,2438E-03	-3,9008E-03
Mean diagonal length	1,2616E-02	1,0422E-01	-1,3931E-02	5,9189E-02
Recurrence period density	4,0592E + 00	5,9616E + 00	1,3806E + 00	-4,4946E + 00
Recurrence rate	6,6090E + 00	1,0217E + 01	-6,5513E + 00	8,8718E + 00
Recurrence time1	-1,5261E + 00	1,6415E + 00	-5,5278E + 00	9,3753E-01
Recurrence time2	-5,3167E-03	-4,8862E-03	2,3380E-02	5,0237E-02
Transitivity	-1,7820E + 01	-1,1284E + 01	1,6588E + 01	2,1919E + 00
Trapping time	2,4517E-03	-8,2911E-02	-4,2520E-02	-1,3478E-01
Proportion of Trace	0.6066	0.1977	0.1317	0.0281

Loading coefficients of the linear discriminants RQA features

**Table 12** LDA Plankton images

STANDARD & RQA	LD1	LD2	LD3	LD4
Area	8,3468E-06	-8,5546E-06	2,3240E-05	1,2083E-05
Compact	2,8489E + 00	-1,1584E + 00	-6,0993E-01	1,1657E + 01
Contrast	-4,3517E-06	4,8338E-07	-1,6700E-05	-4,2536E-06
Eccentricity	-3,3070E + 00	-5,1458E-01	-2,8622E-01	-1,9799E + 00
HU1	8,1832E + 02	-3,9860E + 02	1,5677E + 02	-7,3236E + 00
Homogeneity	-1,0228E + 01	1,1102E + 01	6,8220E + 00	-7,5339E-01
Length boundary	-2,0340E-04	2,1637E-04	6,8411E-04	2,4935E-04
Solidity	1,7137E + 00	1,7562E + 00	4,4218E + 00	-7,7258E + 00
Clustering coefficient	6,8073E + 00	-1,8136E + 01	-4,4149E-01	-1,8076E + 01
Determinism	-6,0420E + 01	-6,0060E + 01	8,9124E + 00	7,5204E + 01
Entropy Diagonal length	-1,2037E-01	6,6271E-01	2,2168E + 00	8,3095E-01
Laminarity	8,1272E + 01	6,4581E + 01	-5,1721E + 01	-1,3652E + 02
Longest diagonal length	-2,0340E-04	2,1637E-04	6,8411E-04	2,4935E-04
Longest vertical length	4,4576E-04	-5,3818E-04	-5,0247E-03	4,8416E-04
Mean diagonal length	-7,4275E-04	-3,6588E-02	-6,5053E-02	-2,0883E-02
Recurrence period density	-6,5591E-01	-4,1900E + 00	-3,3693E + 00	3,1853E + 00
Recurrence rate	-3,1627E + 00	-9,4214E + 00	1,2058E + 00	3,9744E + 00
Recurrence time1	1,7024E + 00	-4,0523E-01	8,3799E-01	-1,7335E + 00
Recurrence time2	-1,2678E-02	1,3795E-02	-1,1402E-02	-2,7078E-04
Transitivity	3,2961E + 00	1,7547E + 01	-6,0033E + 00	6,6110E + 00
Trapping time	1,9478E-02	8,2368E-03	8,0562E-02	1,2844E-02
Proportion of Trace	0.4158	0.3135	0.1500	0.0578

Loading coefficients of the linear discriminants using the STANDARD &amp; RQA features

**Table 13** LDA Plankton images

STANDARD, RQA & ESQ	LD1	LD2	LD3	LD4
Area	7,9128E-06	-7,4814E-06	-2,0536E-05	-1,9393E-05
Compact	3,5193E + 00	-1,9512E + 00	5,9329E-01	-1,1089E + 01
Contrast	-4,0052E-06	-2,7805E-06	1,3528E-05	7,4554E-06
Eccentricity	-3,0145E + 00	-1,2020E + 00	1,3014E-01	1,6067E + 00
HU1	8,3916E + 02	-3,0765E + 02	-8,7893E + 01	-7,9133E + 01
Homogeneity	-1,0727E + 01	7,8797E + 00	-8,9586E + 00	7,7979E-01
Length boundary	-4,6620E-04	-6,1266E-04	-1,4366E-03	1,3844E-04
Solidity	1,1912E + 00	2,3733E + 00	-4,3375E + 00	6,8645E + 00
Clustering coefficient	7,3574E + 00	-1,4573E + 01	3,1754E + 00	1,7005E + 01
Determinism	-6,7257E + 01	-5,9122E + 01	-6,8447E + 00	-6,9468E + 01
Entropy diagonal length	-1,0037E-01	9,8246E-01	-1,8145E + 00	-1,1907E + 00
Laminarity	9,4881E + 01	5,1073E + 01	4,1259E + 01	1,2378E + 02
Longest diagonal length	-4,6620E-04	-6,1266E-04	-1,4366E-03	1,3844E-04
Longest vertical length	6,7974E-04	-4,3464E-04	5,0085E-03	-1,0819E-04
Mean diagonal length	4,3436E-03	-4,0438E-02	6,0747E-02	3,6780E-02
Recurrence period density	-1,3176E-01	-4,2263E + 00	3,7156E + 00	-3,2918E + 00
Recurrence rate	-3,2024E + 00	-5,5922E + 00	2,4596E + 00	-7,2368E + 00
Recurrence time1	1,6874E + 00	7,3116E-01	2,1852E-02	8,4096E-01
Recurrence time2	-1,1802E-02	1,0745E-02	7,7042E-03	9,9184E-03
Transitivity	2,8574E + 00	1,4969E + 01	2,9577E + 00	-5,0812E + 00
Trapping time	1,6490E-02	1,3611E-02	-7,2016E-02	-3,7084E-02
Mean size eyes	-5,7648E-06	1,3136E-05	8,6921E-06	-2,4460E-05
Median pixels in eyes	5,2174E-06	-9,6114E-06	-3,9947E-06	1,2096E-05
Num eyes	5,3709E-03	-2,6717E-03	2,0588E-03	-5,2218E-03
Sum pixels in eyes	1,4634E-07	6,8848E-07	6,0735E-07	-1,2721E-07
Proportion of Trace	0.4015	0.3218	0.1488	0.0600

Loading coefficients of the linear discriminants using the *STANDARD*, *RQA* and *ESQ* features

The multivariate analyses revealed that neither *RQA* nor *RQA & ESQ* are well suited as exclusive features for the classification task at hand. Nevertheless, used in combination with the *STANDARD* features, they increased discrimination success.

An important feature of the *STANDARD* feature class was the HU1-moment, which is scale and transformation invariant. Therefore it is able to describe the characteristic shape of an organism irrespective to camera rotation in plane view or magnification. One of the key features of the *RQA* was the *Recurrence rate*, which simply gives the density of observed recurrences indicating the degree to which the organism's contour line exhibits repetitions of similar structures (e.g. polychaete parapodia). It is thus a measure of the structural regularity of the organism. The key *RQA* features *Laminarity* and *Determinism* focus on the vertical and diagonal structures. These two features have been shown before to be some of the most characteristic properties of an RP (details on *RQA* and how to read an RP can

be found in [20]). They characterise diagonals and vertical lines and thus, length and type of contour line segment similarity. The key feature *Transitivity* further gives a probability on the phase space neighbourhood situation.

As successive roots explain less of the observed variance, the general discrimination success is often identified by plotting the scores of the first roots (Figs. 2 and 3). Within the plots better clustering of objects of the same class and better separation among classes mean a higher discrimination success. The ability to discriminate between classes of similar shape structure can be improved by using *RQA* parameters. There is also an indication, that the use of *ESQ* can further improve discrimination between objects of different size classes and regularity (e.g. Appendicularians and Vertebrata vs. Crustaceans), but does not improve general classification. However, these improvements can be used to separate at least 1–2 classes from the entire population. After excluding identified classes a downstream model with less classes allows improving discrimination during the next iterations.

**Table 14** LDA Geometric shapes

Analyses	Hyper-plane side	LD1	LD2	LD3	LD4
STANDARD	-	- Compact - Eccentricity	- HU1 - Solidity	- Compact - HU1	- Compact - HU1
	+	- HU1 - Homogeneity	- Compact - Homogeneity	- Solidity	- Solidity
RQA	-	- Laminarity - Transitivity	- Clustering coefficient - Determinism	- Clustering coefficient	- Clustering coefficient - Laminarity - Recurrence rate
	+	- Determinism	- Laminarity - Transitivity	- Recurrence time 1 - Determinism - Recurrence Rate - Transitivity	- Determinism - Recurrence period density - Transitivity
STANDARD & RQA	-	- Homogeneity - Clustering coefficient - Laminarity	- Determinism - Compact - Homogeneity - Clustering coefficient	- Solidity - Recurrence rate - Clustering coefficient - Laminarity	- Recurrence rate - Transitivity - Laminarity - Recurrence time 1
	+	- Eccentricity - Determinism - Compact - HU1 - Recurrence rate - Transitivity - Recurrence Time 1	- HU1 - Solidity - Recurrence rate - Transitivity - Laminarity	- Determinism - Compact - HU1 - Transitivity	- Determinism
STANDARD & RQA & ESQ	-	- HU1 - Homogeneity - Clustering coefficient - Laminarity	- Determinism - Compact - Homogeneity - Clustering coefficient	- Solidity - Recurrence rate - Clustering coefficient - Laminarity	- Recurrence rate - Transitivity - Laminarity - Recurrence Time 1
	+	- Eccentricity - Determinism - Recurrence rate - Transitivity - Recurrence time 1	- HU1 - Solidity - Recurrence rate - Transitivity - Laminarity	- Determinism - Compact - HU1 - Transitivity	- Determinism - HU1

Importance of features included in LDA. Features were considered to be important when their loading reached at least 10 % of the maximum loading on the respective side of the hyperplane, set up by the discriminant roots (LD)

### Study design

This study is a first conceptual approach to introduce and test the general usability of RQA and ESQ feature sets for image classification. In the overview presented here, some pre-tests and verifications (e.g. [24]) have been intentionally neglected and the approach was directly applied to a highly diverse plankton set. Criticisms may include, that objects were analysed by using a ‘one-fits-all’ embedding approach and analysed diagonal/vertical line length histograms for several features included lengths as low as 2 recurrence points. Nevertheless, it was found that classificatory systems can benefit from the use of RQA features. Thus, this paper primarily sketches out the method and gives first examples

how to use it. We assume that the ESQ features gain higher importance with decreasing neighbourhood threshold  $\varepsilon$ . Thus, future work needs to focus on avoidance of potential problems and consideration of specific adaptations. In detail it seems appropriate to use recurrence analyses with RQA and ESQ specifically in tailored models, to first split distinct classes from the image population. In succeeding steps then better customised RQA and ESQ with adjusted values for  $m$ ,  $t$  and  $e$  can be used.

It is also obvious that some of the included features, especially those that characterise textural properties, are barely sufficient for proper discrimination of the geometrical line art shapes. Respectively the parameter *Contrast*

**Table 15** LDA Plankton images

Analyses	Hyper-plane side	LD1	LD2	LD3	LD4
STANDARD	-	- Eccentricity - Homogeneity	- Homogeneity	- Homogeneity	- HU1
	+	- HU1	- HU1	- Compact - HU1	- Compact
RQA	-	- Laminarity - Transitivity	- Entropy Diagonal length - Transitivity	- Clustering coefficient - Laminarity	- Laminarity
	+	- Determinism	- Clustering coefficient - Laminarity - Recurrence period density - Determinism - Recurrence rate	- Determinism - Transitivity	- Determinism - Recurrence rate
STANDARD & RQA	-	- Determinism - Homogeneity	- Determinism - HU1	- Transitivity - Laminarity	- Clustering coefficient - Laminarity
	+	- HU1	- Homogeneity - Transitivity - Laminarity	- HU1	- Determinism - Compact
STANDARD & RQA & ESQ	-	- Determinism - Homogeneity	- Determinism - HU1	- - HU1 - Homogeneity	- Determinism - HU1 - Compact
	+	- HU1 - Laminarity	- Transitivity - Laminarity - Homogeneity	- Laminarity	- Clustering coefficient - Laminarity

Importance of features included in LDA. Features were considered to be important when their loading reached at least 10 % of the maximum loading on the respective side of the hyperplane, defined by the respective root

showed negligible loadings (Table 4, Tables 6 and 7). However, to date these features are important in automated plankton discrimination and often appear to be among the most important ones in plankton discrimination [12].

It is also clear that Linear Discriminant Analysis is not the most powerful classificatory system available for such multivariate data. As an LDA tries to insert separating hyperplanes in a dimensional space that is defined by the number of given variables, linear classifications often fail. Especially for low inter- and high intra-class variances, as generally expected for in-situ plankton images, it is recommended to apply methods for mapping input features into higher dimensional spaces, using the kernel trick (e.g. Support Vector Machines). However, the advantage of a LDA is the simple access and interpretation of the feature loadings and thus an initial assessment of the importance of the different variables.

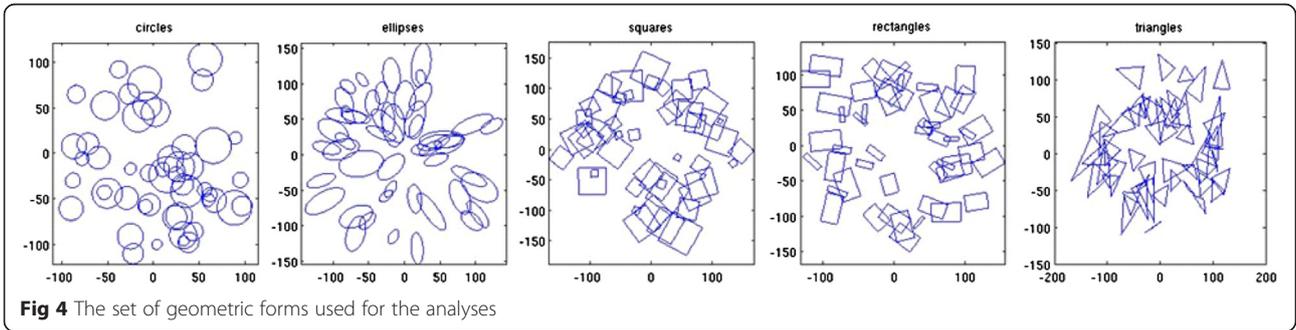
## Conclusions

It could be shown, that the principle of recurrence plots and subsequent analyses can be applied to contour line data of imaged and pre-segmented objects. The tailored embedding algorithm enabled our application to derive

new image features for automated classification systems of plankton organisms. Additionally, a new set of features was derived by measurement of contiguous elements of given phase space dissimilarity (eye-structures in the recurrence plots).

The discriminative success of the LDA was enhanced by using a combination of standard image features, recurrence quantification analysis features and the newly proposed eye-size features. This improvement was observed both for the synthetic data set of geometric and the real-world phytoplankton images. The characterization of images by recurrence quantification analysis and eye structure quantification offers auxiliary image features that could not be derived by applying standard image features alone. We recommend the use of the standard features in combination with the features derived from the application of recurrence analysis to discriminate between classes of phytoplankton. With further improvements the class of such methods may further improve automated plankton identification, which represents an important step forward in the effective processing of large numbers of under-water images and autonomous monitoring stations.

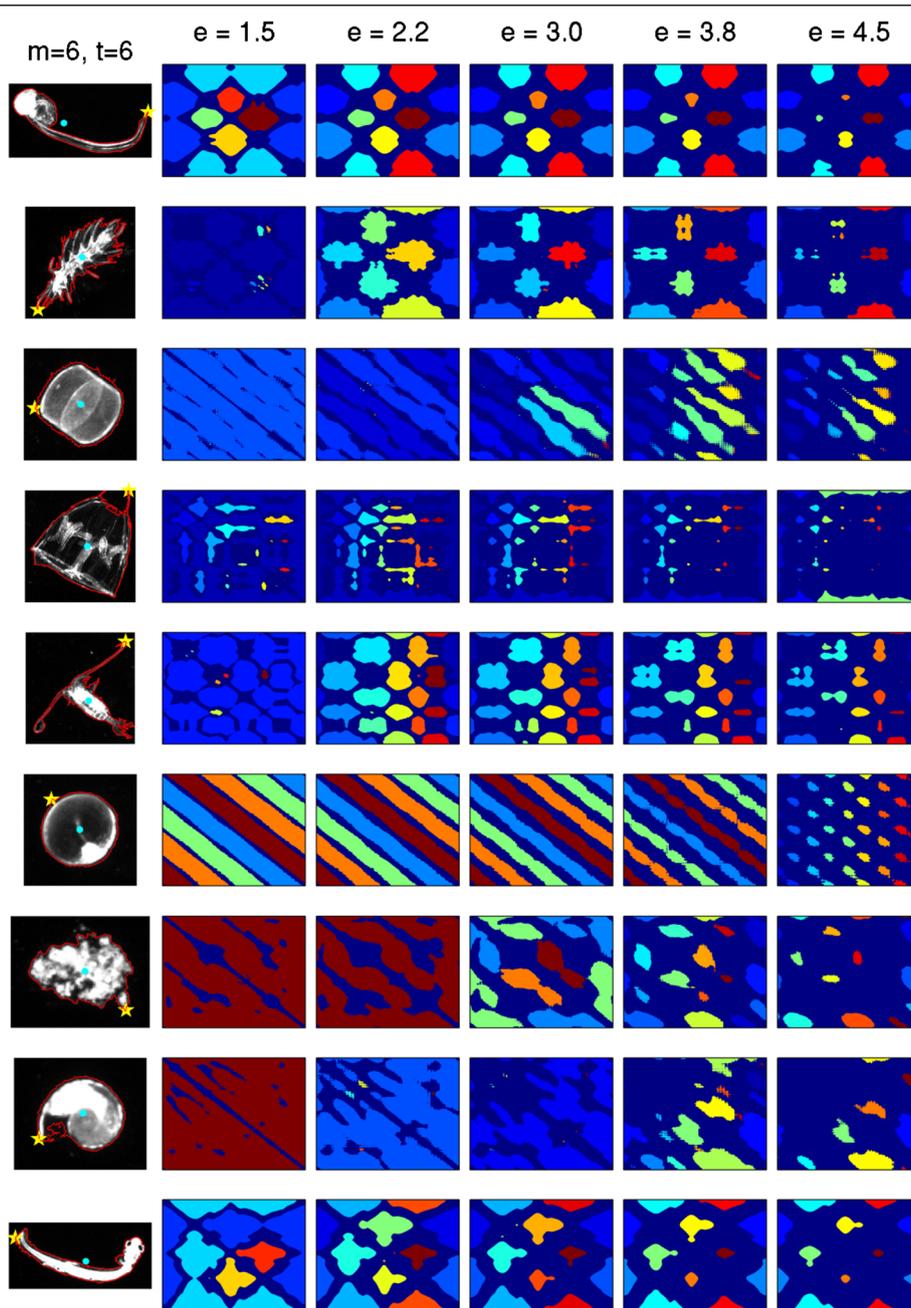
## Appendix A



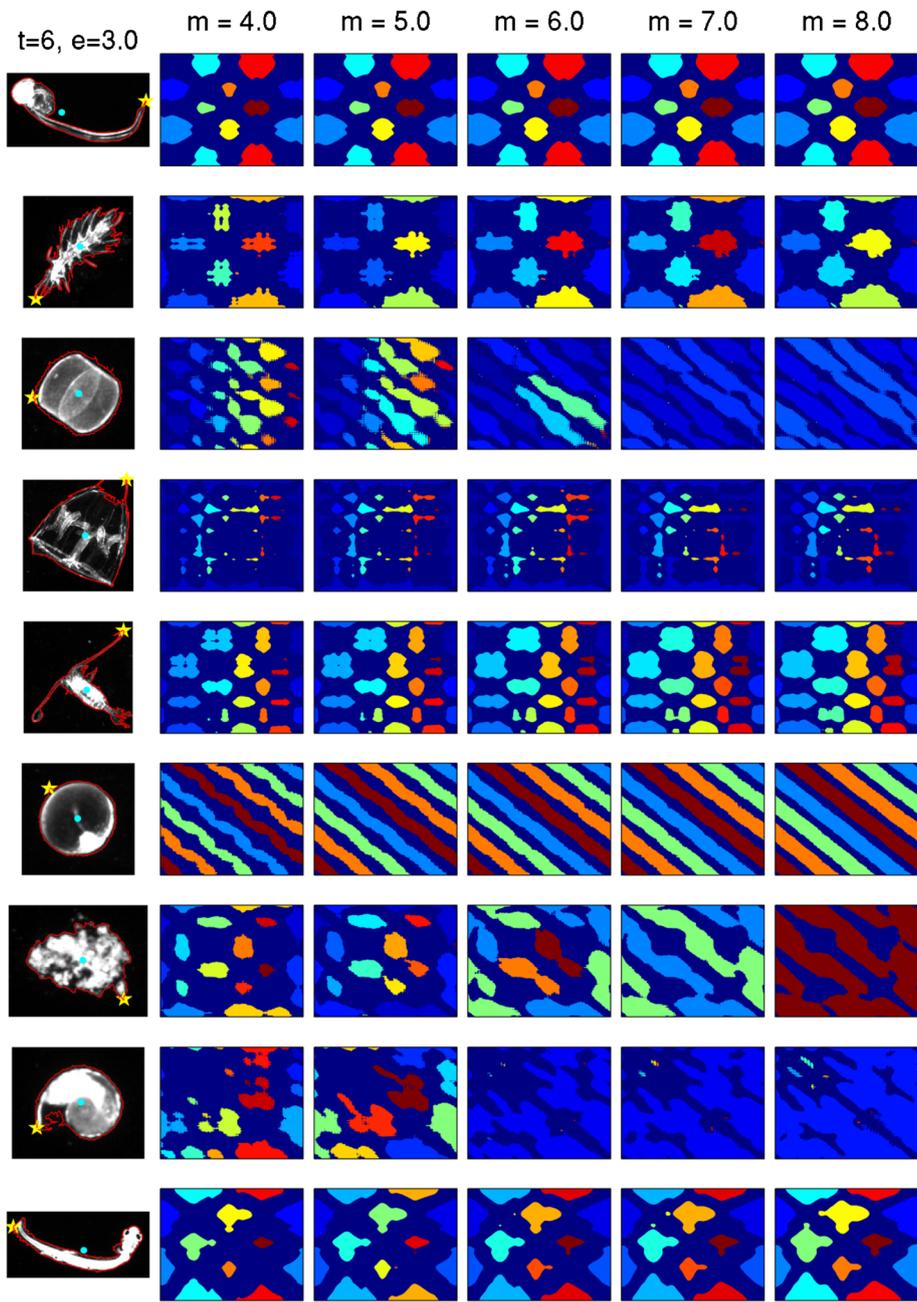
**Appendix B**

In the following some recurrence plot panels are shown for different values of  $m$ ,  $t$  and  $\epsilon$ . To the left one sample organism is shown for each of the 9 taxonomic/morphologic classes (Appendicularia, Annelida, Bacillariophyceae, Cnidaria, Crustacea, Dinoflagellata, Marine snow, Mollusca, and Vertebrata). The red line

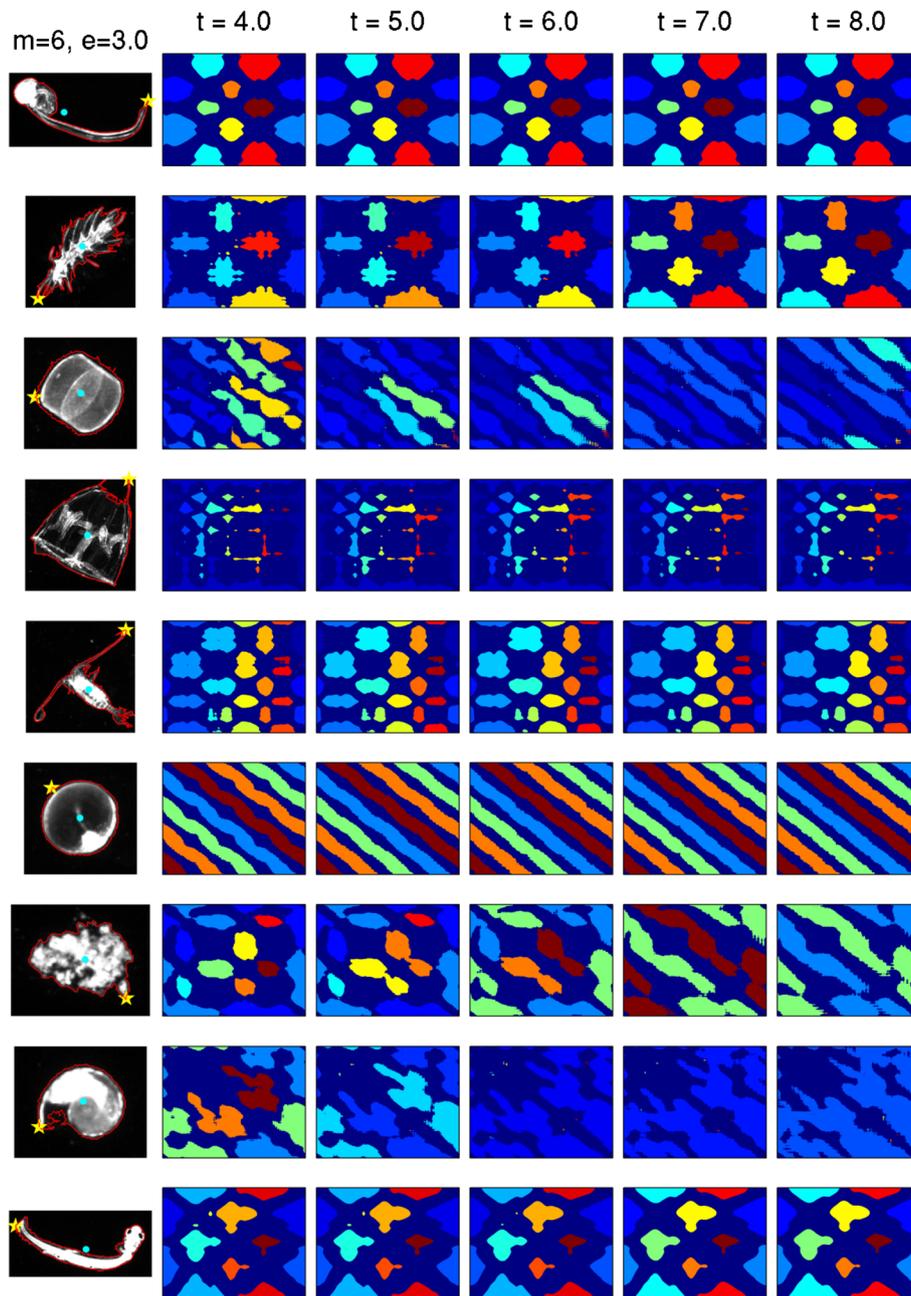
marks the extracted organism’s contour line. The cyan dot inside the imaged object area indicates the object’s centroid. The yellow star displays the first entry of the contour line, which is the contour line point with the highest distance to the centroid. The following columns show the recurrence plots for varying recurrence parameters.



**Fig. 5** Recurrence plots for  $m = 6$ ,  $t = 6$  and varying  $\epsilon$



**Fig. 6** Recurrence plots for  $t = 6, \epsilon = 3$  and varying  $m$



**Fig. 7** Recurrence plots for  $m = 6$ ,  $\varepsilon = 3$  and varying  $t$

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

JS: Idea, preparation of the manuscript, general programming and statistical evaluation in R. AM: Code implementation in Matlab and R, general analysis, preparation of data sets and general contribution to the manuscript. OZ: General contribution to the manuscript. All authors read and approved the final manuscript.

Received: 30 September 2015 Accepted: 30 March 2016  
Published online: 23 June 2016

#### References

1. Eckmann, J.-P., Kamphorst, S.O., Ruelle, D.: Recurrence Plots of Dynamical Systems. *Europhys. Lett.* **4**, 973–977 (1987)
2. F. Takens, "Detecting strange attractors in turbulence" in *Lecture Notes in Mathematics*, David Rand, and Lai-Sang Young, ed., 366–381 (Springer, Warwick, 1981). doi://10-1007-BFb0091924.
3. J. P. Zbilut and C. L. Webber, "Embeddings and delays as derived from quantification of recurrence plots" *Physics Letters A* **171**, 199–203 (1992). doi://10.1016/0375-9601(92)90426-M.
4. Webber, C.L., Zbilut, J.P.: Dynamical assessment of physiological systems and states using recurrence plot strategies". *J. Appl. Physiol.* **76**, 965–973 (1994)

5. N. Marwan, N. Wessel, U. Meyerfeldt, A. Schirdewan, and J. Kurths "Recurrence-plot-based measures of complexity and their application to heart-rate-variability data" *Physical Review E* 66, 026702 (2002). doi://10.1103/PhysRevE.66.026702.
6. Schulz, J., Barz, K., Ayon, P., Lüdtke, A., Zielinski, O., Mengedoht, D., Hirche, H.-J.: Imaging of plankton specimens with the Lightframe On-sight Keyspecies Investigation (LOKI) system". *J. Eur. Opt. Soc. Rapid Publ* 5, 10017s (2010)
7. MacLeod, N., Benfield, M.C., Culverhouse, P.F.: Time to automate identification. *Nature* 467, 154–155 (2010)
8. Persoon, E., Fu, K.S.: Shape discrimination using Fourier descriptors. *IEEE Trans. Syst. Man Cybern.* 7, 170–179 (1977)
9. Mokhtarian, F.: Silhouette-Based Isolated Object Recognition through Curvature Scale Space. *IEEE Trans. Pattern Anal. Mach. Intell.* 17, 539–544 (1995)
10. D.G. Lowe "Object recognition from local scale-invariant features" *Proceedings of the International Conference on Computer Vision 2*, 1150–1157 (1999). doi://10.1109/ICCV.1999.790410.
11. Bay, H., Tuytelaars, T., Van Gool, L.: SURF – Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Axel, P. (eds.) *Computer Vision – ECCV 2006*, pp. 404–417. Springer Verlag, Berlin Heidelberg (2006)
12. Hu, Q.: "Application of statistical learning theory to plankton image analysis" PhD Thesis Massachusetts Institute of Technology and Woods Hole Oceanographic Institution, Supervisors: Cabell S. Davis and Hanumant Singh. (2006)
13. J. Schulz, K. Barz, P. Ayon, and H.-J. Hirche "A sample data set of plankton and particles for automated image classification systems sampled off the Peruvian coast" *Pangaea Data Publisher System for Earth & Environmental Science*, Registration in progress.
14. H.-J. Hirche, K. Barz, P. Ayón, and J. Schulz "High resolution vertical distribution of the copepod *Calanus chilensis* in relation to the shallow oxygen minimum zone off northern Peru using LOKI, a new plankton imaging system" *Deep Sea Research Part I* 88, 63–73 (2014). doi://10.1016/j.dsr.2014.03.00.
15. Patton, D.R.: A Diversity Index for Quantifying Habitat Edge. *Wildl. Soc. Bull.* 3, 171–173 (1975)
16. Hu, M.K.: Visual Pattern Recognition by Moment Invariants. *IRE Trans. Inf. Theory* IT-8, 179–187 (1962)
17. Gonzalez, R.C., Woods, R.E., Eddins, S.L.: "Script: invmoments" *Digital Image Processing Using MATLAB*, Revision: 1.5, Date: 2003/11/21 14:39:19, Prentice-Hall. (2004)
18. Marwan, N. "Cross Recurrence Plot Toolbox for Matlab, Reference Manual. Version 5.17, Release 28.16" <http://tocsy.pikpotsdam.de/CRPtoolbox/>.
19. J. Gao, and H. Cai, "On the structures and quantification of recurrence plots". *Physical. Letters, A* 270, 75–87, doi://10.1016/S0375-9601(00)00304-2
20. Webber, C.L., Marwan, N.: *Recurrence Quantification Analysis* (Springer International Publishing. (2015)
21. Fisher, R.A.: The utilization of multiple measurements in taxonomic problems. *Ann. Eugenics* 7, 179–188 (1936)
22. Jennrich, R.I.: Stepwise regression. In: Enslein, K., Ralston, A., Wilf, H.S. (eds.) *Statistical Methods for Digital Computers*. Wiley, New York (1977)
23. Jennrich, R.I.: Stepwise discriminant analysis". In: Enslein, K., Ralston, A., Wilf, H.S. (eds.) *Statistical Methods for Digital Computers*. Wiley, New York (1977)
24. Marwan, N.: How to avoid potential pitfalls in recurrence plot based data analysis. *Int. J. Bifurcation Chaos* 21, 1003–1017 (2011)

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](http://springeropen.com)

---