ISEIS

Journal of Environmental Informatics 33(1) 47-55 (2019)

Journal of Environmental Informatics

www.iseis.org/jei

A Biological Monitoring Method based on the Response Behavior of *Caenorhabditis Elegans* to Chemicals in Water

I. S. Jeong[1], S. R. Lee[2], I. Song[3], and S. H. Kang[4 *]

[1]*Division of Bio-Medical Informatics, Center for Genome Science, National Institute of Health, Korea Centers for Disease Control and Prevention, 187 Osongsaengmyeong2-ro, Cheongju-si, Chungcheongbuk-do 28159, Korea*
[2]*Department of Information and Electronics Engineering, Mokpo National University, 1666 Yeongsan-ro, Muan-gun, Jeonnam 58554, Korea*
[3]*School of Electrical Engineering, Korea Advanced Institute of Science and Technology, 291 Daehag Ro, Daejeon 34141, Korea*
[4]*Department of Information Security, Dongshin University, 185 Geonjae-ro, Naju, Jeonnam 58245, Korea*

Received September 06 2015; revised June 17 2016; accepted August 02 2016; published online June 16 2017

ABSTRACT. In this paper, based on the behavior of *Caenorhabditis elegans* (*C. elegans*) in response to a toxic substance, we propose a novel biological monitoring method for the detection of water contamination. Both before and after the introduction of formaldehyde into the water at the concentration of 0.1 ppm, the swimming activities of *C. elegans* are continuously recorded by a charge coupled device camera at the rate of four frames per second. The behavior in each of the image frames is characterized by the branch length similarity (BLS) entropy profile. The shapes quantified by the BLS entropy profiles are classified into seven shape patterns via the self-organizing map combined with the *k*-means clustering algorithm. Subsequently, a monitoring scheme composed of two hidden Markov models decides the water quality based on the sequence of shape patterns over a certain observation time. The performance of the proposed method is generally affected by the observation interval; yet, experimental results show an accuracy of about 83% for an observation time of five minutes. It is also observed that, by taking the distribution of individual decisions into account, the accuracy of the proposed method can be improved up to 93% and the false negative rate can be reduced to 10%.

Keywords: biological monitoring method, Caenorhabditis elegans, branch length similarity entropy, water management, machine learning

1. Introduction

With advances in modern industry and agriculture, various pollutants have leaked into water bodies, causing serious water pollution. An accurate and efficient monitoring method has emerged as an essential factor in the effective management of water quality and aquatic ecosystems. For real-time monitoring, sensor-based methods are widely used by detecting changes in the physicochemical factors such as pH, dissolved oxygen demand, and biochemical oxygen demand. However, use of devices for such methods requires expensive analysis and manpower (Gunatilaka, 2001).

In the meantime, a wide range of methods to assess water quality using indicator species have been studied, from the molecular level to communities and ecosystems (Bae, 2014). Among the various methods, monitoring techniques based on the behavior of an organism seem to be the most effective means of linking small and large scale assessments (Bae, 2014). Burridge et al. (2000) studied the response behavior of sea louse (a parasite on lobsters) to a short-term exposure to azamethiphos and cypermethrin. Roast et al. (2000) examined behavioral disruption, particularly of swimming ability, in the hyperbenthic mysid *N. intger* by the effects of chlorpyrifos, an organophosphate pesticide. These studies mainly focused on the effect of a specific pesticide used to get rid of parasites in fishery products to maintain salability. Results from these research on response behavior at the individual level are applied in the development of biomonitoring systems to determine whether pollutants have been introduced or not.

Shedd et al. (2001) proposed an automated biomonitoring method using the differences in the ventilation frequency, whole body movement, ventilator depth, and cough frequency of the bluegill (*Lepomis macrochirus*) before and after exposure to toxic substances. Based on the response behavior of seabream (*Sparus aurata*) and turbot (*Scophthalmus maximus*) at an acute hypoxic test condition (2 mg O_2 l^{-1}), an online biomonitor system was presented to monitor the quality of marine water in real time (Cunha, 2008). Other biomonitoring schemes includes those utilizing abnormal swimming behavior of *Daphnia* (Jeon, 2008) and valve-gape of mussels (Kramer, 19 89). These biomonitoring methods are based commonly on signals such as ventilator frequency and cough frequency collected from electrodes in water, which usually contain noise,

* Corresponding author. Tel.: +82-61-330-3953; fax: +82-61-330-3029.
E-mail address: kinston@gmail.com (S. H. Kang).

ISSN: 1726-2135 print/1684-8799 online
 doi: 10.3808/jei.201700356

and consequently, require elaborate signal processing techniques.

Recently, with advances in image processing techniques and related digital equipments, the response behavior of indicator species has been explored by utilizing digital images in many studies. Park et al. (2005) suggested a method based on image processing techniques for detecting and analyzing the response behavior of medaka (*Oryzias Latipes*). To characterize the response behavior, the method employs several measures such as speed, angle, and angular speed calculated from the movement tracks of medaka exposed to diazinon, as recorded by a charge coupled device (CCD) camera. Liu et al. (2011) proposed a method to analyze behavioral changes of zebrafish (*Danio rerio*) in response to formaldehyde.

These methods, commonly based on digital image processing techniques, showed the possibility of improving the accuracy of monitoring systems and determining water quality in real time. Unfortunately, these methods provide only supportive information about the water quality, but not automatic alarms. In other words, the final decision on whether a body of water is contaminated with pollutants or not is entirely dependent on human experts, not on the system itself.

The main contribution of this paper is to demonstrate that *C. elegans* can be used as a bio-indicator in the monitoring of water quality and to propose a novel method for the monitoring of water quality using *C. elegans* as a bio-indicator. Recently, with the help of image processing techniques, the swimming behavior of *C. elegans* has been shown to follow a Markov process (Kang, 2012a). The possibility of the swimming behavior of *C. elegans* as a measure for biological early warning systems was briefly discussed as well (Choi, 2012; Kang, 2012a). The proposed method is based on the behavioral change of *C. elegans* after the introduction of formaldehyde into the water at a concentration of 0.1 ppm. The method also employs digital image processing and machine learning techniques such as the self-organizing map (SOM) and hidden Markov model (HMM). Additionally, taking the distribution of individual decisions over a certain period into consideration, a strategy for improving the performance of the proposed method is also suggested and tested.

2. Materials and Methods

2.1. Organisms and Experimental Set-up

In this study, 40 adult individuals of wild type N2 *C. elegans* were considered. They were cultivated in Petri dishes(60 mm in diameter and 15 mm in height) filled with growth medium for nematode in an incubator at 20 °C and fed with *Escherichia coli* of strain OP50.

We made small circular arenas with a diameter of 4.0 mm and a depth of 2.0 mm. The arenas for the control group (composed of 20 randomly selected specimens) were filled with deionized distilled water without any chemical treatment. For the treated or control group (composed of the remaining 20 specimens), formaldehyde at a concentration of 0.1 ppm was directly added to the water in the arena.

Formaldehyde is a colorless and strong-smelling gas, often found in water-based solutions. In addition to being an industrial pesticide, it is commonly used in the production of particle board, household products, and paper coatings. In the light of its widespread use and toxicity, formaldehyde is a significant factor for human health and is even known to be a human carcinogen (Shaham, 1996). It has been reported that a 1% concentration of formaldehyde is lethal to nematodes (Moerman, 1981), and that airborne concentrations above 0.1 ppm can cause irritation of the eyes and respiratory tract in human adults (OSHA, 2011). Although the species is differrent, the behavioral changes of zebrafish under 0.1 ppm concentration of formaldehyde have been studied (Liu, 2011). We chose the concentration of 0.1 ppm based on these observations.

Immediately after introducing the organism into the arena, a cover glass was placed on the arena to prevent water evaporation. The specimen was acclimated for about 10 min. The swimming behavior of each specimen was monitored and recorded at four frames per second for 40 minutes with a CCD camera during the day under the natural light at the same temperature of 20 °C as in the incubator. The swimming behavior for each of the 40 nematodes was individually observed, producing 9,600 frame images in total from each specimen.

2.2. Image Processing

As a preprocessing phase to characterize the behavior of *C. elegans*, 13 points at equal intervals along the length of the organism in each image were extracted using image processing techniques (Gonzalez, 2002). Firstly, the background image (Figure 1(a)) was subtracted from the target image with a nematode at a certain position (Figure 1(b)). Then the image was transformed into a binary image containing the organism alone by averaging and appropriate thresholding for noise removal. Finally, by segmenting the binary image and applying a skeletonization algorithm, 13 points were placed at equal intervals along the length of the individual.

2.3. Branch Length Similarity Entropy

In this study, we used the branch length similarity (BLS) entropy (Lee, 2010a, b) to characterize the swim- ming behavior of *C. elegans*. The BLS entropy has been successfully used in a number of shape recognition schemes for identifying objects such as the human face (Lee, 2011) and butterflies (Kang, 2012b, 2014). The BLS entropy is defined on a simple branch network, referred to as the unit branch network (UBN), consisting of a single node and its branches. The BLS entropy S is defined as:

$$S = -\frac{1}{\log(n)} \sum_{i=1}^{n} p_i \log(p_i) \quad (1)$$

where

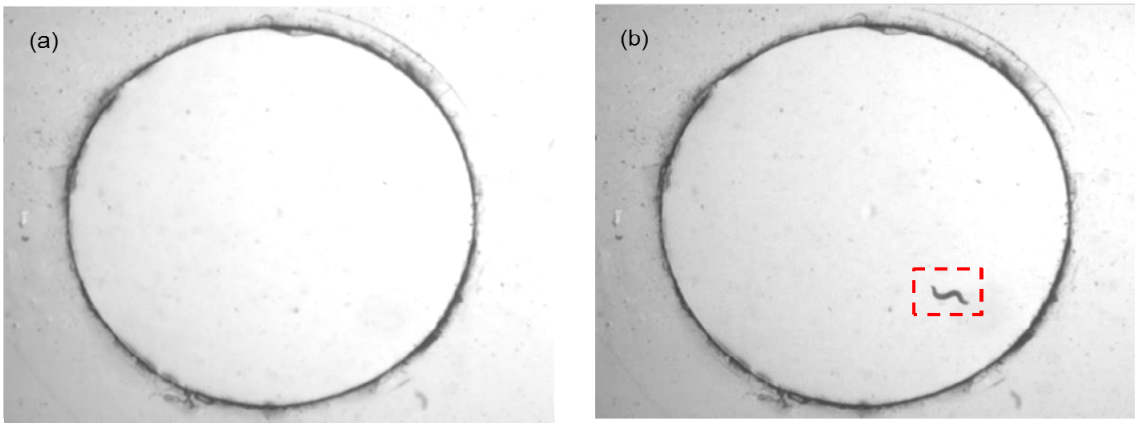

Figure 1. A nematode object is taken out by subtracting the background image from the target image. a) background image; b) target image.

$$p_i = \frac{l_i}{\sum_{k=1}^{n} l_k}$$

with l_i representing the length of the i-th branch and n the number of branches in the UBN.

The BLS entropy for the *C. elegans* in an image is computed by applying Equation (1) to the network composed of the 13 points placed at equal intervals along the length of an individual. Figure 2 shows an example of the UBN to calculate the BLS entropy S_7 for point 7. The assembly (S_1, S_2, …, S_{13}) of BLS entropy for the 13 consecutive points is called the BLS entropy profile, and is used as a descriptor to characterize the shape of *C. elegans* in an image.

2.4. Representation of Swimming Behavior

Owing to the large number of elements and the precision of each element, the BLS entropy profile can be used as an input feature to a machine learning method such as the hidden Markov model (HMM) only after an appropriate partitioning into several patterns. To address this problem, we propose a partitioning procedure composed of the self-organizing map (SOM) (Kohonen, 1982, 2001) and the k-means clustering algorithm (Hartigan, 1975).

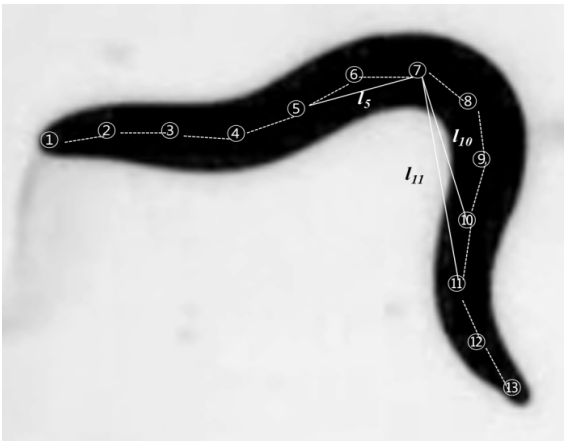

Figure 2. An example of a UBN at node 7 with 12 branches.

The SOM is one of the representative clustering techniques dividing data into clusters through unsupervised learning. Owing to its usefulness and relative simplicity, the SOM has been widely used in various studies such as the data analysis of environment monitoring (Li, 2015; Riga, 2015) and ecological modelling (Bae, 2014) as well as data visualization (Oyana, 2009).

The SOM maps a high dimensional input data into a two-dimensional grid map while preserving the topological properties of the input data. The basic structure of the SOM consists of the input and neuron layers, where the neurons are arranged in a two-dimensional grid and each neuron is directly connected to neighbor neurons and to every node in the input layer.

In the proposed method, the SOM was trained with BLS entropy profiles iteratively. A 14 × 10 hexagonal lattice was adopted as the structure of the neuron layer of the SOM as it yields a good performance. Before the BLS entropy profiles were fed to the input layer of the SOM for training, the values $\{S_1, S_2, \ldots, S_{13}\}$ of the elements of a BLS entropy profile were linearly scaled to range from 0 to 1. At each training step, a BLS entropy profile was randomly drawn from the profile set. Given a BLS entropy profile S = (S_1, S_2, …, S_{13}), the best matching unit (BMU), a neuron with the prototype vector closest to the input vector S in terms of Euclidean distance, is selected. The prototype vector of the BMU and its topological neighbors are updated and moved closer to the given input vector. After an adequate number of iterations, the neuron layer is spatially organized according to the topological structure of the input dataset.

Although the SOM preserves the topological structure of the input data on a grid map, it does not have the ability to automatically determine the number of shape patterns (Kohonen, 2001). In order to represent the swimming behavior of *C. elegans* with several meaningful shape patterns and to effectively utilize the information provided by the SOM, additional methods are required. For the grouping of the neurons in the

trained SOM into several clusters, we employed the *k*-means clustering algorithm together with Davies-Bouldin (DB) index (Davies, 1979). The *k*-means clustering algorithm partitions map units (neurons in the neuron layer) into clusters so that the error function:

$$E = \sum_{i=1}^{k} \sum_{u \in Q_i} \| u - c_i \|^2 \quad (2)$$

is minimized, where k is the number of clusters, Q_i is the i-th cluster of neurons, and c_i indicates the center of Q_i. Here, 'center' denotes the mean value of the weight vectors of neurons in a cluster. In the *k*-means clustering algorithm, the number k of clusters is not determined automatically but should be given in advance by the user. In order to avoid this arbitrariness, we employ the DB index defined as:

$$DB(k) = \frac{1}{k} \sum_{c=1}^{k} \max_{j \neq i} \left[\frac{D_c(Q_i) + D_c(Q_j)}{D_c(Q_i, Q_j)} \right] \quad (3)$$

where $D_c(Q_i)$ and $D_c(Q_i,Q_j)$ denote the average of the distance between a unit and the cluster center in cluster Q_i and the distance between the centers of Q_i and Q_j, respectively. Input vectors (BLS entropy profiles) associated with the neurons belonging to the same cluster after neuron clustering are considered to have the same shape pattern.

A series of test trials were carried out to determine the parameter value k. The *k*-means clustering is repeated by varying k from 2 to $\lfloor\sqrt{14\times10}\rfloor = 11$ to acquire the best clustering. The value which minimizes the DB index is chosen as the number of clusters.

The procedures for the clustring of shape patterns mentioned above can be summarized as follows:

Firstly, the BLS entropy profile for each image is computed by Equation (1). Each element of the BLS enropy profile is then scaled into a value between 0 and 1. Subsequently, the SOM is trained against the dataset of the scaled BLS entropy profile. The best clustering is selected as an output of the shape pattern clustering by applying the k-means clustering algorithm to the neurons of the trained SOM with the goal of minimizing the DB index.

As a result of the clustering procedure of shape patterns, seven shape patterns were identified. The shapes of specimens represented by the BLS entropy profile were classified into one of the seven shape patterns.

Figure 3 shows the seven classes of shape patterns, which we will call *flat*, *medium flat*, *low flat*, *curve*, *low circular*, *medium circular*, and *circular* according the angles between the lines connecting adjacent points for the 13 points. In sequel, the swimming behavior of *C. elegans* over a time period is characterized by a series of shape patterns, each from the patterns at intervals of 0.25 seconds. The movement behavior of an organism for five seconds (corresponding to 20 frames) would be characterized by a vector $p = (p_1, p_2, \ldots, p_{20})$ of shape patterns, where p_t indicates the shape pattern of the t-th frame.

2.5. Decision Maker

The decision-maker in the proposed method is based on two HMMs, one learned with the normal state data generated from specimens under the 'normal' (unpolluted, clear) water condition and the other with the abnormal state data under the 'abnormal' (polluted, contaminated) water condition.

The HMM is one of the representative machine learning techniques, which usually deals with sequential data with temporal properties. It has been successfully applied to a number of areas such as speech recognition (Juang, 1991), online handwriting recognition (Igarza, 2003), spam mail detection (Gordillo, 2007), and gesture recognition (Wilson, 2001) as well as animal behavior modeling (Bagniewska, 2013).

An HMM is defined by the 5-tuple (N, M, A, B, π) (Rabiner, 1989; Won, et al, 2010), where N is the number of states, M is the number of observation symbols, A is the probability distribution of the state transition indicating the probability of transitioning between states, B is the probability distribution of the observation symbol in a state, and π is the probability distribution of the initial state.

In the proposed method, we have $M = 7$, the number of classes of shape patterns. The number N of states was tuned to bestfit the training data through many experiments. The probability distributions A, B, and π are determined from the learning with training data. We adopted an ergodic (or fully connected between state nodes) model (Rabiner, 1989) as the structure for the two HMMs, and employed the well-known Baum-Welch algorithm (Rabiner, 1989) as the learning algorithm for the two HMMs.

The individual decision (whether the water is polluted or not) for each test shape sequence was made by comparing the log likelihoods of the sequence produced by the two HMMs. If the log likelihood issued by the HMM learned with the normal training data is greater than that issued by the HMM learned with the abnormal training data, the water quality at that instant is decided to be normal. Otherwise, the water quality is decided to be abnormal. This can be described by the

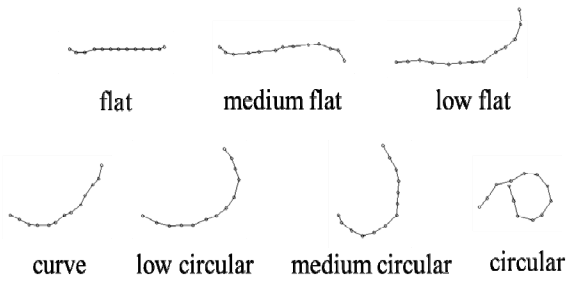

Figure 3. The seven representative shape patterns (*flat*, *medium flat*, *low flat*, *curve*, *low circular*, *medium circular*, *circular*) of *C. elegans* obtained using the SOM combined with the *k*-means clustering algorithm.

individual decision function:

$$w(p)=\begin{cases}0 & (normal),\ if \quad h_N(p)>h_A(p),\\ 1 & (abnormal),if \quad h_N(p)<h_A(p),\end{cases} \tag{4}$$

where p denotes a vector (sequence) of shape patterns, and h_N and h_A are the log likelihoods issued by the HMM learned with the normal and abnormal data, respectively.

The individual decision about the water quality with a shape sequence is simple and efficient; yet a considerable number of miss classifications are produced especially against the abnormal data set. In other words, the method often decided incorrectly the water state as normal when it was in fact in an abnormal state (the experimental results will be presented in the section Results and Discussion). In an effort to resolve this problem and make the scheme more practical, we designed a decision making scheme which provided a final decision by considering the distribution of individual decisions about the water quality over a time period.

Denote a series of shape sequences over a time period as $P = (p^1, p^2, \ldots, p^n)$ where, p^i indicates the i-th sequence of shape patterns. In the proposed method, composed of two trained HMMs, the i-th individual decision about the water quality is made at intervals of given time length using p^i. Over an observation time, the final decision about the water quality is then made based on the ratio of individual positive decisions (warning signals) in that period. If the ratio of individual positive decisions becomes higher than or equal to a pre-defined threshold, the final decision will be that the water is in the abnormal state. Otherwise the water is decided to be in the normal state. The final decision Ω about the water quality over a certain time period can be formulated as:

$$\Omega(P)=\begin{cases}1(abnormal), & if \sum_{k=1}^{\left\lfloor \frac{q}{d} \right\rfloor} w(p^k) \geq \left\lfloor \frac{q}{d}\cdot r \right\rfloor \\ 0(normal), & otherwise,\end{cases} \tag{5}$$

where q indicates the length of the time period over which the final decision is made, d is the length of the interval for individual decisions, and r is the threshold ratio of individual warning signals. For example, assume that an individual decision about the water quality is made at intervals of three minutes (d = 3) over a period of ten minutes (q = 10) with the threshold ratio of 0.7 (r = 0.7). Then, when the number of individual warning signals is greater than $\lfloor 10/3 \times 0.7 \rfloor = 2$, the abnormal state is issued by the proposed scheme as the final decision about the water quality.

3. Results and Discussion

3.1. Performance with Individual Decision

In order to evaluate the performance of the proposed monitoring method, we conducted a number of experiments with various values of the parameters. Firstly, datasets of the sequence of shape patterns were built with observation times ranging from 15 seconds (shape sequence of length 60) to 480 seconds (shape sequence of length 1920) to evaluate the effect of observation length on the performance. Each sequence is generated to overlap with the next sequence by 30% to reduce the gap effect between two consecutive sequences to a certain degree. Table 1 shows the composition of datasets from a total of 40 organisms (20 under normal condition and the other 20 under abnormal condition).

A 20-fold cross-validation was used (Mitchell, 1997) to justify the performance evaluation of the proposed method. A total of 20 rounds of cross-validation were performed using various partitions of specimens, and the validation results were averaged over the rounds. In general, the accuracy of the warning method is measured by:

$$\text{accuracy} = \frac{\text{the number of true positives}}{\text{the total number of test instances}} + \frac{\text{the number of true negatives}}{\text{the total number of test instances}} \tag{6}$$

where a *true positive* occurs when the method correctly detects the abnormal state (water pollution), and likewise, a *true negative* is produced when the method correctly declares the normal state.

Figure 4 plots the accuracies measured over the validation data. Each experiment was performed with HMMs with various numbers (from six to nine) of hidden states and sequences of various length (from 15 to 480 seconds). It is observed that HMMs composed of nine hidden states yield the best performance. In addition, the proposed method achieved an accuracy of 82.9 ± 12.9 % when the input data with an observation time of 480 seconds was used. The results also indicate that, in order to guarantee an accuracy above 80%, at least 960 frames (corresponding to 240 seconds of observation time) are required before the method can make a decision about the water quality.

As an effort to investigate detailed aspects of the performance, we examined the performance with respect to the true positive and true negative rates separately. Figure 5 plots the

Table 1. The Composition of the Data Set

Observation time (sec)	Length of an instance (frames)	The number of instances
15	60	9120
30	120	4520
60	240	2240
120	480	1120
240	960	520
300	1200	440
480	1920	240

*generated from 40 adult individuals of wild type *C. elegans*, 20 under normal water condition and 20 under abnormal water condition.

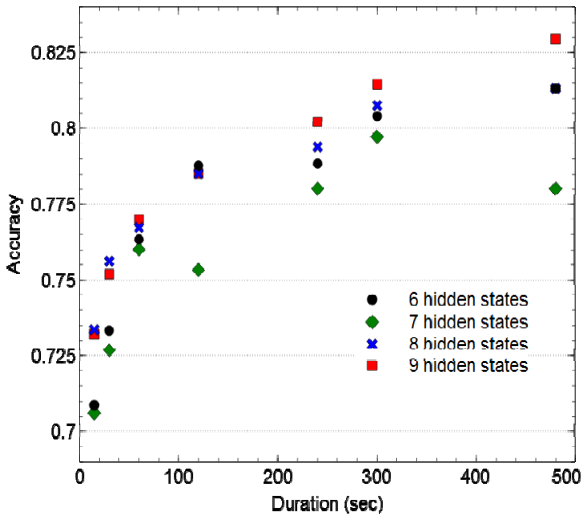

Figure 4. The accuracy of the warning system based on individual decision over the test data with observation time varying from 30 to 480 seconds. Experiments were carried out by varying the number of hidden states of the HMM from six to nine.

changes in the true positive rate (the lower line in the figure) and the true negative rate (the upper line in the figure) along with the average accuracy when an HMM with nine hidden states is employed.

The true negative rate over the test data with observation times of not less than 60 seconds shows an accuracy of above 95%, whereas the true positive rate over the same test data is lower than 70%. This result might have come from the fact that the swimming behavior of *C. elegans* after exposure to a toxic substance often, but not always, shows the same temporal patterns as it was before the exposure (Anderson, 2004; Kang, 2012a). Clearly, although a chemical may stimulate the neuro system of an organism, resulting in the characteristic swimming behavior of the chemically treated condition (Kang, 2012a), the chemical does not dominate the swimming pattern all the time. Therefore, it is not unexpected that organisms exposed to a toxic substance show sometimes the swimming pattern of the normal water condition.

Unfortunately, this fact makes it difficult to decide the water quality correctly at a certain moment solely by the behavior (sequence of shape patterns) of organisms, which we believe is a common vulnerability in every bio-monitoring method.

3.2. Performance with Combined Decision

In order to make our scheme more useful, we have taken into account the distribution of individual decisions over a time period in making the final decision. Firstly, original data from the 40 organisms (normal and abnormal states) were divided into segments (each corresponds to a decision period) of 10 minutes. From these newly generated data sets, 6 segments (three each from the normal and abnormal state sets) were randomly drawn and then combined as shown in Figure 6. Ten test instances were generated for experiments in this manner, resulting in a total of 60 decision periods. The values of ratio r were selected in the range of 5 to 20%. For observation periods ranging from 15 to 480 seconds, two types of experiments were conducted with individual decision intervals of 5 and 10 seconds.

Figure 6 shows two examples of individual decisions at intervals of 10 seconds for a one hour test composed of alternating 10-minute normal and abnormal periods. Figures 6(a) and 6(b) show the distributions of individual decisions using shape sequences of 60 frames (corresponding to an observation time of 15 seconds) and 240 frames (corresponding to an observation time of 60 seconds), respectively. Since the interval of an individual decision is set to 10 seconds, there exist 60 individual decisions in each of the 10-minute time periods. From the figure, it is observed that the modified method correctly decides the water states of 11 periods using Equation (6) except for the last one in Figure 6(b), where all of the 60 individual decisions are 'normal' while the water is in fact abnormal.

In conducting the test, we have taken into consideration that the procedure to process images and to calculate BLS entropy profiles for a sequence sample is expected to require at least 5 seconds or longer. In addition, when a toxic chemical is detected in the water body, to provide a proper information before incurring significant damage, we have adopted a period of 10 minutes for a decision. Of course, these choices can be modified as appropriate.

Figure 7 shows the average accuracy of the proposed scheme with the distribution of individual decisions (at intervals of 5 and 10 seconds) into account, in which the averaging was performed over several threshold rates from 5 to 20%. The method decided the water quality of 60 periods

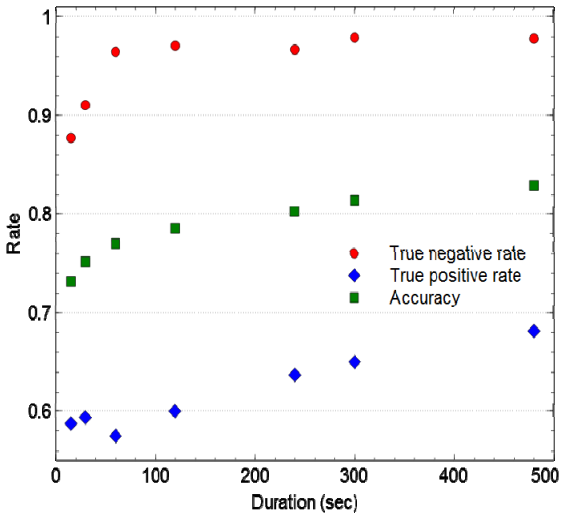

Figure 5. The true positive and true negative rates of the proposed method based on individual decision with nine hidden states.

(comprising of 30 normal and 30 abnormal periods) using Equation (6). Except in one case of observation time of 15 seconds, the accuracy with a 5-second interval decision was higher than that with a 10-second interval. When the threshold ratio, observation length, and individual decision interval are 10%, 480 seconds, and 5 seconds, respectively, the method showed an accuracy above 93%.

It is also observed from the experiments that the true positive and true negative rates have tradeoff relations with respect to the threshold: If the threshold ratio decreases, then the true positive rate increases whereas the true negative rate decreases. This tradeoff relation was conspicuous especially in the case of a short observation time (e.g. 15 and 30 seconds). Figure 8 shows the tradeoff between the true positive and true negative rates when the individual decisions were made at intervals of five seconds with observation time of 30 seconds.

These results imply that we should choose the parameters according to the requirements of the water environment. For example, when a significant damage is expected even with a small amount of pollutants, a lower threshold ratio and a shorter individual decision time interval are recommended.

3.3. Discussion

We have so far demonstrated that the swimming behavior of *C. elegans* combined with digital image processing techniques and machine learning methods could be successfully used to monitor the water quality. In order to better understand the locomotion behavior, researchers have focused on patterns of movements of *C. elegans* and their distributions, and also on how to characterize them. For example, Choi et al. (Choi, 2012) showed that the swimming behavior of nematodes could be classified into several shape patterns by using the distributions of distances and angles between points along the length of body. In addition, they observed that the distributions of patterns were different between before and after chemical treatments. Kang et al. (Kang, 2010) extended these experiments and found that the temporal sequence of patterns followed a Markov model. In addition, they showed that the temporal pattern sequences were different between before and after chemical treatments by using the distribution of pattern sequences and Levenshtein distance, one of the metrics to measure the similarity between sequences. In these studies, the possibility of nematode as a bio-indicator is also briefly discussed. In essence, we have in this study confirmed the possibility of using the response behavior of nematode to chemicals as a bio-indicator suggested previously.

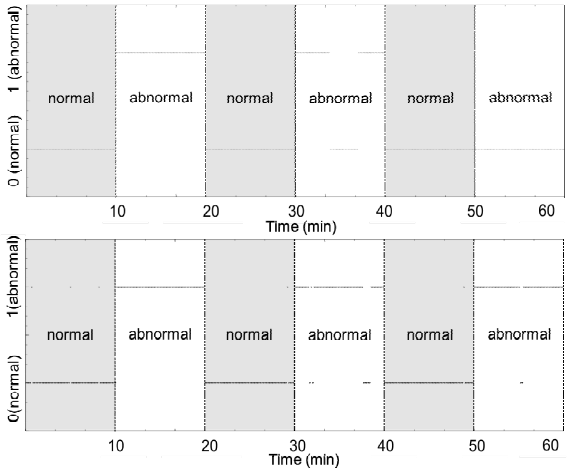

Figure 6. Two examples showing the individual decisions about the water quality at intervals of 10 seconds for a one hour test composed of alternating 10-minute normal and abnormal periods. a) Result of an experiment using shape sequences of 60 frames (corresponding to observation time of 15 seconds); b) Result of an experiment using shape sequences of 240 frames (corresponding to observation time of 60 seconds).

The proposed method could be extended to other slender bio-indicators such as earthworm. Interestingly, earthworm has attracted many researchers as a bio-indicator in the evaluation of soil conditions (Paoletti, 1999, Rombke, 2005); the conventional methods are mainly based on abundance or biomass. We believe that our method, based on the individual behavior of earthworm in response to specific chemicals, could be more useful.

In spite of the successful results described above, the chemical used to bring about the response of *C. elegans* in our scheme was limited to formaldehyde. It is necessary to carry out more research on the response behavior of organisms to other chemicals such as benzene and toluene. In addition, the investigation of how to measure the changes in the behavior of an organism at different levels of concentration of chemicals is also an interesting research topic for the development of a more complete biological monitoring and early warning systems.

4. Conclusions

In this study, we have proposed a novel bio-monitoring scheme based on the responses of *C. elegans* to formaldehyde. The study has demonstrated the possibility of nematode as a successful bio-indicator in the assessment of water quality when an appropriate scheme is employed for pattern recognition and decision making. The BLS entropy profile, a series of the BLS entropy for 13 points along the length of *C. elegans*, was employed as the main feature to characterize the behavior of nematodes. The SOM combined with the k-means clustering algorithm was used to partition the BLS entropy profile and the swimming behavior of *C. elegans* was characterized by a sequence of seven shape patterns. In determining the water quality, the HMM, a representative machine learning method, was employed.

With an observation time of five minutes, the results from experiments exhibited an initial accuracy of about 83%. To enhance the applicability of the proposed method, the ratio of individual positive decisions over a 10-minute period was exploited to finally decide the water quality over the period.

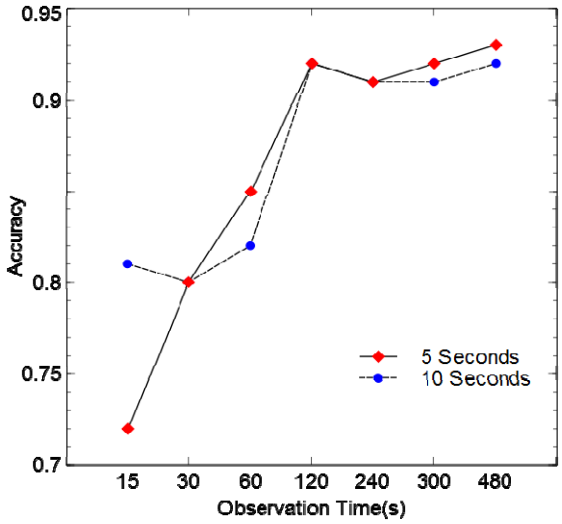

Figure 7. The average accuracy of the proposed scheme with the distribution of individual decisions (at intervals of 5 and 10 seconds) into account, in which the averaging was performed over several threshold rates from 5 to 20%.

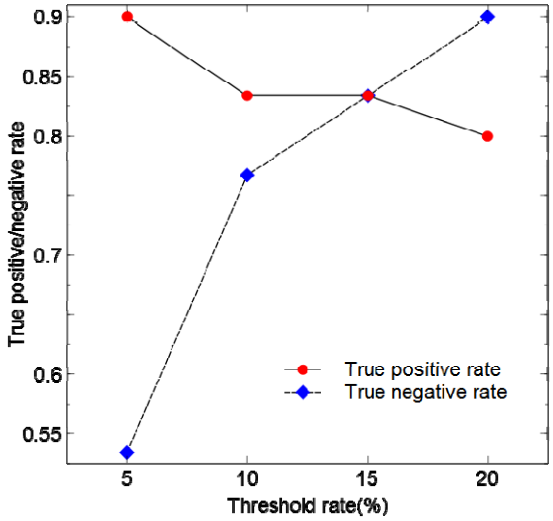

Figure 8. Variations of the true positive rate and the true negative rate of the proposed scheme against threshold rates, where individual decision was made at intervals of five seconds over an observation time of 30 seconds.

The results of this modified method showed an accuracy of above 94% while maintaining a high true positive rate. In short, the proposed method based on the behavior of *C. elegans* is expected to help the development of biological monitoring and early warning systems with other organisms also.

Acknowledgments. The authors wish to thank the Editor and three anonymous reviewers for their invaluably constructive suggestions and helpful comments. This work was supported by the National Research Foundation of Korea under Grant NRF-2014R1A1A4A01008-799 with funding from the Ministry of Education and under Grant NRF-2015R1A2A1A01005868 with funding from the Ministry of Science, Information and Communications Technology, and Future Planning.

References

Anderson, G.L., Cole, R.D., and Williams, P.L. (2004). Assessing behavioral toxicity with Caenorhabditis elegans. *Environ. Toxicol. Chem., 23(5),* 1235-1240. http://dx.doi.org/10.1897/0 3-264

Bae, M.J., and Park, Y.S. (2014). Biological early warning system based on the response of aquatic organisms to disturbances: A review. *Sci. Total Environ.,* 466-467, 635-649. http://dx.doi.org/10.1016/j.scitotenv.2013.07.075

Bagniewska, J.M., Hart, T., Harrington, L.A., and Macdonald, D.W. (2013). Hidden Markov analysis describe dive patterns in semi-aquatic animals. *Behav. Ecol.,* 24(3), 659-667. http://dx. doi.org/10.1093/beheco/ars217

Burridge, L.E., Haya, K., Waddy, S.L., and Wade, J. (2000). The lethality of anti-sea lice formulations Salmosan® (Azamethiphos) and Excis® (Cypermethrin) to stage IV and adult lobsters (Homarus americanus) during repeated short-term exposures. *Aquaculture,* 182(1-2), 27-35. http://dx.doi.org/10.1016/S0044-8486(99)00251-3

Choi, Y.T., Jeon, W., Kang, S.H., and Lee, S.H. (2012). Characterizing temporal patterns in the swimming activity of Caenorhabditis elegans. *J. Korean Phys. Soc.,* 60(11), 1840-1844. http://dx.doi. org/10.3938/jkps.60.1840

Cunha, S.R., Goncalves, R., Silva, S.R., and Correia, A.D. (2008). An automated marine biomonitoring system for assessing water quality in real-time. *Ecotoxicol.,* 17(6), 558-564. http://dx.doi.org/10.1007/s10646-008-0216-y

Davies, D.L., and Bouldin D.W. (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.,* 1(2), 224-227. http://dx.doi.org/10.1109/TPAMI.1979.4766909

Gonzalez, R.C., and Woods, R.E. (2002). *Digital Image Procesing,* Second edition, Prentice Hall.

Gordillo, J., and Conde, E. (2007). An HMM for detecting spam mail. *Expert Syst. Appl.,* 33(3), 667-682. http://dx.doi.org/10.1016/j.eswa.2006.06.016

Gunatilaka, A., and Diehl, P. (2001). A Brief Review of Chemical and Biological Continuous Monitoring of Rivers in Europe and Asia. (in *Biomonitors and Biomarkers as Indicators of Environmental Change,* Vol. 2), New York, Kluwer Academic, 9-28. http://dx.doi.org/10.1007/978-1-4615-1305-6_2

Hartigan, J.A. (1975). *Clustering algorithms,* John Wiley & Sons.

Igarza, J.J., Goirizelaia, I., Espinosa, K., Hernaez, I., Mendez, R., and Sanchez, J. (2003). Online handwritten signature verification using hidden Markov models. *Lecture Notes in Computer Science,* 2905, 391-399. http://dx.doi.org/10.1007/978-3-540- 24586-5_48

Jeon, J., Kim, J.H., Lee, B.C., and Kim, S.D. (2008). Development of a new biomonitoring method to detect the abnormal activity of Daphnia magna using automated Grid Counter device. *Sci. Total Environ.,* 389(2), 545-556. http://dx.doi.org/10.1016/j.scitotenv.2007.09.015

Juang, B.H., and Rabiner, L.R., (1991). Hidden Markov models for speech recognition. *Technometrics,* 33(3), 251-272. http://dx.doi.org/10.1080/00401706.1991.10484833

Kang, S.H., Cho, J.H., and Lee, S.H. (2014). Identification of butterfly based on their shapes when viewed from different angles using an artificial neural network. *J. Asia-Pac. Entomol.,* 17(2), 143-149. http://dx.doi.org/10.1016/j.aspen.20 13.12.004

Kang, S.H., Chon, T.S., and Lee, S.H. (2012a). Exploring the behavior of Caenorhabditis elegans by using self-organizing map

and hidden markov model. *J. Korean Phys. Soc.,* 60(4), 604-612. http://dx.doi.org/10.3938/jkps.60.604

Kang, S.H., Jeon, W., and Lee, S.H. (2012b). Butterfly species identification by branch length similarity entropy. *J. Asia-Pac. Entomol.,* 15(3), 437-441. http://dx.doi.org/10.1016/j.aspen.2012.05.005

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biol. Cybern.,* 43(1), 59-69. http://dx.doi.org/10.1007/BF00337288

Kohonen, T. (2001). *Self-Organizing Maps,* Springer Berlin Heidelberg. http://dx.doi.org/ 10.1007/978-3-642-97966-8

Kramer, K.J.M., Jenner, H.A., and Zwart, D.D. (1989). The valve movement response of mussels: a tool in biological monitoring. *Hydrobiologia,* 188(1), 433-443. http://dx.doi.org/10.1007/BF00027811

Lee, S.H. (2010a). Robustness of the branch length similarity entropy approach for noise added shape recognition. *J. Korean Phys. Soc.,* 57(3), 501-505. http://dx.doi.org/10.3938/jkps.57. 501

Lee, S.H., Bardunias, P., and Su, N.Y. (2010b). A novel approach to shape recognition using the shape outline. *J. Korean Phys. Soc.,* 56(31), 1016-1019. http://dx.doi.org/10.3938/jkps.56.101 6

Lee, S.H., Kim, E.Y., and Yi, D. (2011). Characterizing facial expressions in males and females by using branch length similarity entropy. *J. Korean Phys. Soc.,* 58(2), 377-380. http://dx.doi.org/10.3938/jkps.58.377

Li, W., Zhang, H.T., Zhu, Y., Liang, Z.W., He, B., Hashmi, M.Z., Chen, Z.L., and Wang, Y.S. (2015). Spatiotemporal classification analysis of long-term environmental monitoring data in the northern part of lake taihu, china by using a self-organi zing map. *J. Envirnon. Inf.,* 26(1), 71-79. http://dx.doi. org/10.4.2014/JEN.In

Liu, Y., Lee, S.H., and Chon, T.S. (2011). Analysis of behavioral changes of zebrafish (Danio rerio) in response to formaldehyde using self-organizing map and a hidden Markov model. *Ecol. Medel.,* 222(14), 2191-2201. http://dx.doi.org/10.1016/j.ecolmodel.2011.02.010

Mitchell, T.M. (1997). *Machine Learning,* McGraw-Hill.

Moerman, D.G., and Baillie, D.L. (1981). Formaldehyde mutagenesis in the nematode Caenorbabditis elegans. *Mutation Res.,* 80(2), 273-279. http://dx.doi.org/10.1016/0027-5107(81) 90100-7

OSHA (2011). *Formaldehyde,* Occupational Safety and Health Administration. https://www.osha.gov/OshDoc/data_General_Facts/formaldehyde-factsheet.pdf.

Oyana, T.J. (2009). Visualization of high-dimensional clinically acquired geographic data using the self-organizing maps. *J. Environ. Inf.,* 13(1), 33-44. http://dx.doi.org/10.3808/jei.20090 0138

Paoletti M.G. (1999). The role of earthworms for assessment of sustainnability and as bioindicators. *Agr. Ecosyst. Environ.,* 74 (1), 137-155. http://dx.doi.org/10.1016/S0167-8809(99)00034- 1

Park, Y.S., Chung, N.I., Choi, K.H., Cha, E.Y., Lee, S.K., and Chon, T.S. (2005). Computational characterization of behavioral response of medaka (Oryzias latipes) treated with diazinon. *Aquat. Toxicol.,* 71(3), 215-228. http://dx.doi.org/10.1016/j.aquatox.2004. 11.002

Rabiner, L.R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE,* 77(2), 257-286. http://dx.doi.org/10.1109/5.18626

Riga, M., Stocker, M., Ronkko, M., Karatzas, K., and Kolehmainen, M. (2015). Atmospheric environment and quality of life information extraction from twitter with the use of self-organizing map. *J. Environ. Inf.,* 26(1), 27-40. http://dx.doi.org/10.3808/jei.201500311

Roast, S.D., Widdows, J., and Jones, M.B. (2000). Disruption of swimming in the hyperbenthic mysid Neomysis integer (Peracarida mysidacea) by the organophosphate pesticide chlorpyrifos. *Aquat. Toxicol.,* 47(3), 227-241. http://dx.doi.org/10.1016/S0166-445X(99)00016-8

Rombke, J., Janscha, S., and Didden, W. (2005). The use of earthworms in ecological soil classification and assessment concepts. *Ecotox. Environ. Safe.,* 62(2), 249-265. http://dx.doi.org/10.1016/j.ecoenv.2005.03.027

Shaham J., Bomsten Y., Meltzer, A., Kaufman, Z., Palma, E., and Ribak, J. (1996) DNA-protein crosslinks, a biomarker of exposure to formaldehyde-in vitro and in vivo studies. *Carcinogenesis,* 17(1), 121-125. http://dx.doi.org/10.1093/carcin/17.1.12 1

Shedd, T.R., van der Schalie, W.H., Widder, M.W., Burton, D.T., and Burrows, E.P. (2001). Long-term operation of an automated fish biomonitoring system for continuous effluent acute toxicity surveyllance. *Bull. Environ. Contam. Toxicol.,* 66(3), 392-399. http://dx.doi.org/10.1007/s00128-001-0018-x

Wilson, A.D., and Bobick, A.F. (2001). Hidden Markov Models for Modelling and Recognizing Gesture under Variation. *Int. J. Patt. Recogn. Artif. Intell.,* 15(1), 123-160. http://dx.doi.org/10.1142/S0218001401000812

Won, S.H., Song, I., Lee, S.Y., and Park, C.H. (2010). Identification of finite state automata with a class of recurrent neural networks. *IEEE Tr. Neural Networks,* 21(9), 1408-1421. http://dx.doi.org/10.1109/TNN.2010.2059040