Characterization of alum floc in water treatment by image analysis and modeling

Petri Juntunen, Mika Liukkonen, Markku Lehtola and Yrjö Hiltunen

Cogent Engineering (2014), 1: 944767
Characterization of alum floc in water treatment by image analysis and modeling

Petri Juntunen1*, Mika Liukkonen1, Markku Lehtola1 and Yrjö Hiltunen1

Abstract: In many water treatment plants, flocculation is the key unit concerning the performance of water treatment. For this reason, monitoring the flocculation (i.e. floc size) is a crucial issue to achieve the acceptable performance for the process. Generally, flocculation is monitored by naked eye or using complex, sample-based methods. This is laborious and expensive, however, and should be either automated or alternative methods for estimating the floc quality should be developed if possible. In this paper, we present an online characterization system for estimating the most essential quality parameters of floc using digital images taken in the flocculation unit. In addition, we compare the surface area of the floc particles defined using the images with other measurement data collected from the process, and create a multivariable regression model for it. We also illustrate the dependencies between the floc properties and other process variables using a self-organizing map.

Keywords: water treatment, water quality, image analysis, flocculation, regression, self-organizing map

ABOUT THE AUTHOR

The University of Eastern Finland (UEF) enjoys a leading national position in the field of forestry research. The research projects of the Process Informatics research group in the Department of Environmental Science have focused on advanced measurement and modeling methods and advanced systems utilized by the process industry. Research has been done within the energy, pulp, chemical, electronics, and water industries. So far we have developed and applied intelligent methods, which can be exploited in offline or online software tools. These tools can be used in process monitoring and optimization, for example. Currently, we are developing advanced online monitoring systems in which we combine novel measurement technology and advanced data processing to extract new kind of information from processes. In addition, several commercial software tools have been developed which have been adopted by service business companies operating in the process industry.

PUBLIC INTEREST STATEMENT

In this article, we introduce a low-cost camera application, which is suitable for characterization of alum floc in water treatment processes. The application is capable to produce different floc describing parameters to produce parameter such as floc size or floc number by image analysis. In addition, the calculated parameters are combined with other process parameters and analyzed by data-driven modeling methods such as multiple linear regression or self-organizing maps. The application enables a fruitful way to achieve more online information about the most essential unit of the water treatment process.
1. Introduction

Flocculation is a critical unit process of drinking water treatment. It involves the formation of floc followed by the aggregation of floc that is amenable to solid/liquid separation with subsequent processes such as sedimentation, flotation, and/or filtration. The most common coagulants used in water treatment are alum and other aluminum-based chemicals. However, the flocculation process is extremely complex because many chemical and physical features of raw water affect the aluminum coagulation and flocculation process. Speaking of chemical features, many organic and inorganic compounds in suspended, colloid, or solved form influence the process. Organic compounds such as fulvic or humic acids play an essential role in the coagulation and flocculation. Furthermore, many inorganic compounds such as the $\text{SiO}_2^-$, $\text{OH}^-$, $\text{F}^-$, $\text{PO}_4^{3-}$ or $\text{SO}_4^{2-}$ affect the process. As an example of physical parameters, temperature of the water has a remarkable influence on the flocculation process (Kvech & Edwards, 2002; van Benschoten & Edzwald, 1990). Moreover, it is evident that the process conditions itself have a great effect on Al coagulation and flocculation. The dose of the Al chemical is, of course, one of the key parameters, as is the adjusted pH value. Furthermore, hydraulic variables in flocculation units such as velocity or velocity gradient ($G$-value) affect flocculation. Running variables of the separation processes such as surface load or washing periods influence the performance (Letterman, 1999). Moreover, in water treatment, there are observable cycles and/or episodic events such as rapid changes, which cause the process to behave dynamically (Bratby, 2006; Juntunen, Liukkonen, Pelo, Lehtola, & Hiltunen, 2013). In summary, the coagulation and flocculation processes are both physically and chemically heterogeneous, for which the development of online monitoring applications is extremely challenging, and time-consuming and expensive in situ testing such as z-potential metering may be needed.

For these reasons, there is a need for developing new approaches to monitor and manage water treatment and water quality online and to ensure safe drinking water with reasonable costs. Real-time monitoring, which is characterized by a rapid response time, full compatibility with automation, sufficient sensitivity, high rate of sampling, minimal requirements for skill and training, etc. (Mays, 2004) can provide data that can be used in multiple purposes, such as fault monitoring or process assessment. In water treatment, as in many domains, process monitoring and control relies heavily on accurate and reliable sensor information. While many process variables can be measured continuously using relatively simple and cheap physical sensors, the determination of certain quantities of interest requires costly laboratory analyses that cannot be performed online (Valentin & Denoeux, 2001).

One interesting approach to characterize the floc formation is the image analysis of the forming floc. The main advantage of this approach is that we can measure the most essential features of the floc such as the size and the form of the floc. Previously, there have been some studies of the image analysis in ex situ or in situ considering for water treatment (Chakraborti, Atkinson, & Van Benschoten, 2000; Wang, Lu, Du, Shi, & Wang, 2009), or waste water treatment cases (Perez, Leite, & Coelho, 2006). Nonetheless, these applications are either laboratory or pilot-scale studies not designed for online use, which is useful if the purpose is to study the chemistry of floc formation or validate the results of image analysis. Considering online use in full-scale water treatment plants, such applications would be complex and expensive. On the other hand, the information contained by digital images can be exploited without knowing the exact value of the measured parameters. In this case, we make interpretations of the image analysis by comparing the calculated image characterization variables to measured process variables.

Advanced statistical methods, including artificial neural networks, have been applied successfully to various problems in different industrial fields, such as the pulp and paper industry (Alonso, Negro, Blanco, & San Pío, 2009; Avelin, Jansson, Dotzauer, & Dahlquist, 2009). For example, self-organizing maps (SOM) have proved to be useful and efficient in modeling water quality in general level (Kalteh, Hjorth, & Berndtsson, 2008; Maier, Morgan, & Chow, 2004). A SOM is an unsupervised learning method to analyze various data-sets, including those with missing values.
The application of techniques of artificial intelligence, which can be used for analyzing and visualizing complex multivariable data, can potentially be useful in analyzing data, such as features of water quality parameters (Juntunen, Liukkonen, Pelo, Lehtola, & Hiltunen, 2011) or outliers (Muruzábal & Muñoz, 1997). Typical for a SOM is that the desired solutions or targets are not given and the network intelligently learns to cluster the data by recognizing different patterns (Alhoniemi, Hollmen, Simula, & Vesanto, 1999; Kohonen, 2001). A SOM possesses advantages over other multivariate approaches, because it can handle the nonlinearities in a system; it can be produced using the data without mechanistic knowledge of the system; it can deal with noisy, irregular, or missing data; it can be easily and quickly updated; and it can interpret information from multiple variables or parameters with its excellent visualization capabilities (Hong, Bhamidimarri, & Charleson, 1998; Liukkonen, Havia, Leinonen, & Hiltunen, 2011; Liukkonen, Hiltunen, Hälkkä, & Hiltunen, 2011). In addition, SOM can combine the different data-sets to a single form which is visually and easily understandable (Kohonen, 2001).

As far as we know, characterization of floc using a method capable of operating online in a real water treatment process has not been possible this far. In this paper, we present a low-cost online characterization system for estimating the size and other features of the floc in the flocculation unit. The system consists of an ordinary systems camera, which is automated to capture images over the flocculation pool. The images are then analyzed to calculate characteristics that indicate the size of the floc and other features such as eccentricity and the number of floc particles. After characterizing the floc, multiple linear regression (MLR) and SOM are used to analyze the reasons for changes in floc properties.

2. Materials and methods

2.1. Case process and data

In our case process, the water is first bank filtrated. After filtration, the water is purified with chemical coagulation in a chemical purification process. The aluminum flocculant is dosed to the process for flocculation. In addition, lime is dosed to the process for adjusting the pH in two separate locations (A and B later in the text). After coagulation, the coagulated floc is separated by sedimentation or flotation followed by sand filtration. In the final stage, water is disinfected by chlorination.

The measurement period was organized during 26 May 2011 to 12 June 2011. The period was divided into two parts. In the first part, which consisted of the first 11 days, the raw water was pumped mainly from the wells. In the latter period, the raw water was pumped from the lake. The process was shut down in two week-end periods. The resolution of the process data is 1 h. Thirteen of the process variables were selected manually for further investigation (Table 1).

2.2. Measurement system

A low-cost industrial camera system was installed to learn whether digital images could be used in monitoring the floc properties in the flocculation unit and in characterizing the features of the floc.

Table 1. The process variables used in the analysis

<table>
<thead>
<tr>
<th>Level of the wells A*</th>
<th>Derivative of lime feed A**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of the wells B*</td>
<td>Lime feed B**</td>
</tr>
<tr>
<td>Surface/ground water feed ratio</td>
<td>Derivative of lime feed B**</td>
</tr>
<tr>
<td>Flow from wells A*</td>
<td>pH in flocculation</td>
</tr>
<tr>
<td>Flow from wells B*</td>
<td>pH of treated water</td>
</tr>
<tr>
<td>Total flow to process</td>
<td>Pressure cut in filtration</td>
</tr>
<tr>
<td>Lime feed A**</td>
<td></td>
</tr>
</tbody>
</table>

*A and B in the context of wells refer to two separate well groups in different locations.

**A and B in the context of lime feed refer to two separate dosing points of lime.
The system includes a commercial systems camera (Nikon D3100 + 18–55 mm standard objective) and a mobile measuring and analyzing unit for capturing online images from the process. The systems camera is located over the pool of the flocculation unit. The automatically taken images (every 30 min in this case) were recorded on a measurement server and transferred to a remote PC using a wireless connection. In a more sophisticated application, the images could be forwarded to a control room PC, for example, to be viewed and analyzed. The proposed implementation of the system in a water treatment plant is shown in Figure 1.

2.3. Image data and analysis

Images were taken from the process automatically by the system during one month in the summer of 2011. The size of digital images is 4608 × 3072 pixels. After the measurement period, the system had collected 950 digital images. The resolution of process data is one hour, however, and therefore the floc variables calculated from the digital images were averaged to have the same resolution, and thus the final data-set had 464 rows of data.

There are a large number of image-processing tools, which can be potentially used for preparing digital images for the measurement of the features and structures they reveal (Russ, 2011). Analysis of floc images taken from the process is a relatively difficult problem due to many real-world challenges: the illumination can be poor; floc particles are small and overlap with each other; material is in layers; and so on. Fortunately, computational processing of images makes it possible to solve many of these problems.

Various variables were determined to indicate the size, form, and color of the floc particles. A computer program for analyzing floc was coded on the Matlab (Version 7.10) platform and the image-processing toolbox (Version 7.0). The stages for determining the floc characteristics in a single image are as follows:

1. Upload digital color image to analysis.
2. Convert image to grayscale.
3. Convert grayscale image to binary form (zeros and ones) to detect floc particles using a fixed threshold value (.65).
4. Find components (i.e. floc objects) connected by pixels using the neighborhood of four pixels.
5. Calculate desired object properties based on the pixel data.
The following floc properties were calculated for each image:

1. Average surface area of floc particles
2. The number of floc particles
3. Average equivalent diameter of floc particles: the diameter of a circle with the same area as the region
4. Eccentricity of particles: the eccentricity (ratio of the distance between the foci of an ellipse and its major axis length) of the ellipse that has the same second-moments as the region. The value of the eccentricity is between 0 and 1, where 0 is a circle and 1 is a line segment.
5. Average perimeter of floc particles
6. The average red (R), green (G), and blue (B) levels of the color model in digital images.
7. Color index from the indexed color images

2.4. Modeling methods

2.4.1. Multivariate linear regression

In MLR (Cohen, 1968), the purpose is to model the relationship between two and more explanatory variables and a response variable by fitting a linear equation to observed data samples. In principle, the MLR model with observations and variables is given in Equation 1:

\[ Y = e + a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n \]  

where \( Y \) is the value of the response variable, \( X \) is the value of the predictor (explanatory) variable, \( a_0, \ldots, a_n \) equals the unknown coefficient to be estimated, and \( e \) signifies the uncontrolled factors and experimental errors of the model. In this case, \( Y \) is the average surface area of floc particles, and \( X_n \) are the independent explanatory variables used in the model (see Table 1). The fitting works by minimizing the sum of the squares of the vertical deviations from each data point to the line that fits best for the observed data, which is also called least-squares fitting.

In variable selection based on MLR, sequential forward selection method was used. In this approach, the variables are included in progressively larger subsets, so that the goodness of the model is maximized. To select \( p \) variables from the set \( P \):

1. Search for the variable that gives the best value for the selected criterion.
2. Search for the variable that gives the best value with the variable(s) selected in Stage 1.
3. Repeat Stage 2 until \( p \) variables have been selected.

2.5. Self-organizing map

Multivariate measurement data may be difficult to interpret using standard data processing methodology. Advanced descriptive methods such as SOM (Kohonen, 2001) can be useful in detecting nonlinear relationships between data variables. SOM is an unsupervised neural network methodology, which can be used to transform an \( n \)-dimensional input vector into a lattice (or discrete feature map), which is usually two-dimensional. In SOM, the input vectors sharing common features are projected to the same areal units (i.e. neurons) of the map. Each neuron of the feature map has an \( n \)-dimensional weight vector, which works as a link between the input and output spaces in such a way that input vectors with common characteristics are assigned to the same or neighboring neurons. This will preserve the topological order of the original input data. The lattice of neurons reflects variations in the statistics of the data-set and highlights common characteristics that approximate to the distribution of the data points. Each neuron includes an \( n \)-dimensional reference vector (prototype vector), which describes its common properties. The array of neurons (the map) can be illustrated as a rectangular, hexagonal, or even irregular organization, the size of which can be altered depending on the application; the more neurons, the more details are represented. In view of its
ability to compress information, the SOM is an ideal means of analyzing large data-sets typical of industrial processes.

In summary, training of a SOM advances as follows:

1. Initialize the map.
2. Find a best-matching unit (BMU) for the input vector using Euclidean distance.
3. Move the reference vector of the BMU toward the input vector.
4. Move the reference vectors of the neighboring neurons toward the input vector.
5. Repeat Steps 2–4 for all input vectors in sequence.
6. Repeat Steps 2–5 using a smaller learning rate factor (fine-tuning phase).
7. Find the final BMUs for the input vectors.

The power of a SOM lies in presenting the most characteristic features of a multidimensional data-set in a low-dimensional display. This enables easier detection of multivariate interactions in large data-sets. During the learning phase, the input data vectors are mapped one by one into particular neurons based on the minimal n-dimensional (Euclidean) distances between the input vectors and the reference vectors of the map neurons. Each time a new input vector is located, the reference vectors of the activated neurons are updated, eventually leading to a self-organizing network. In this unsupervised methodology, the SOM can be constructed without previous a priori knowledge (Kohonen, 2001).

In the case study, variance scaling was used for pre-processing the data. A 15 × 15 hexagonal grid was used in the map, and linear initialization, batch training algorithm, and a Gaussian neighborhood function were used in training.

3. Results and discussion

Monitoring the floc properties is an important but difficult issue in water treatment plants, because of process dynamics, seasonal and episodic variations, and complex chemical reactions, which are partly unknown. A low-cost online system for estimating the size and other features of the floc in the flocculation unit of a water treatment plant has been presented. Here, we present the results from analyzing the collected digital images and the corresponding measurement data from the process. The analysis stage consisted of the following phases. First, the digital images were analyzed to get average values for floc properties. Next, the obtained data were coded to a SOM together with the measurement data obtained from the process. In addition, a multivariate regression model was created for the average surface area of floc determined from the images.

The average floc surface areas (in pixels) calculated from digital images and two sample images are presented in Figure 2. As can be seen, the index indicates both short- and long-term changes during the measurement period. In addition, it can be seen that the floc size is obviously smaller when ground water is treated (period 1) than when surface water is used (period 2).

Furthermore, the calculated floc variables are shown in Figures 3 and 4. Again, we can observe a change in the behavior of several variables when the ground water period is changed into the surface water period. Especially, the average red value of the RGB space drops remarkably after this transition.

In Figure 5, the SOM component planes of the process and image data are presented. As can be seen, there seems to be plenty of similarity between the patterns of eccentricity and the derivatives of lime feed A and B. In other words, the form of the floc particles seems to correlate with the lime feed. Furthermore, the lime feed seems to correlate with the calculated surface area of floc. It can...
Figure 2. The calculated average values of the floc area (in pixels) and two samples from the test period.

Figure 3. The ratio of surface water during the test period and the calculated average values of the floc diameter, eccentricity, perimeter, and number (in pixels).
also be seen that the area of floc particles becomes larger, in general, when surface water is treated (see Figure 2).

Next, we present the regression model created for the average surface area of flocs. In Figure 6, the results of variable selection for the average surface area of flocs using process data and multivariate linear regression are shown. The so-called “cumulative” correlation is presented here, which is produced by the variable selection procedure. In other words, the first correlation value is the correlation of a regression model including only the topmost (i.e. the most important) variable, the second value is the correlation of a model including the two most important variables, and so on. It is notable that the first two of the variables selected are lime feed variables, which confirms that there is a connection between the lime feed and the surface area of the floc particles forming in flocculation. The five variables presented produce a relatively good model for the average surface area ($R \approx .82$). This model can be seen in Figure 7. As can be seen, the model is able to follow the baseline of the calculated value.

Validation of the results of image analysis is of course difficult. The only way in practice would be to use naked eye, and calculate the number of particles manually, for example. The other option would be to use a microscope in parallel with the systems camera, calculate the floc properties manually from the microscope images, and see if the results would match with those given by the
approach presented here. This would be extremely difficult in a large scale, however. Therefore, in practice, the only reasonable way to get comparison data for images is to use naked eye (e.g. see Figure 2).

Furthermore, the same method is used for the number and R component (see Figures 8 and 9).

Furthermore, in this case, following conclusions can be made from Figures 6 to 9:

- **Size** and **form** of floc particles: pH important
- **Number** of particles: pH and water source important
- **Color** of floc particles: water source important. For example, the ground water wells were identified as the potential sources of iron, because there is an observable drop in the R value when surface water is taken into use

On the other hand, when monitoring the process we should be more interested in changes occurring in the floc quality, after all, and not so much on the absolute values of its different properties. The quality parameters calculated from digital images offer one possible tool for monitoring these changes, and it can be used (1) for diagnosing the reasons for changes when combined with corresponding process data and (2) as the basis of a warning system, which would alarm when either
sudden or long-term changes occur in the process. The system also has potential in real-time use, because the processing of images and calculation of floc property parameters are not computationally heavy. It would be possible to install the monitoring system into the control room of a water treatment plant, as presented in Figure 1, where an online image would be shown together with a long-term floc property indexes.

Models for floc variables vary from moderate to good; there are clear connections between floc properties and the process. The results show that especially the average surface area and eccentricity of the floc seem to be the most interesting quality parameters in this case. The source of the raw water (i.e. ground or lake water) seems to have an effect on the size of the floc particles. The changes in the surface area can be explained using five variables, two of which are variables describing lime feed. In addition, the results suggest that the eccentricity of the floc particles seems to provide an indicator of the changes in the process. Especially, the lime feed seems to have an influence on the formation or breaking of the floc particles. This kind of information is extremely important for the planning of the pilot tests. When we can select the right parameters and range, resources, such as time and money, can be saved.
Furthermore, the most important issue for the water treatment plant is: what kind of floc is best for the plant? This should be evaluated case-specifically by comparing with quality variables of purified water, but the usual problem is the lack of proper measurement data. However, the camera analysis is a potential way of studying this.

4. Conclusions
A low-cost inspection system for estimating floc properties in the flocculation unit of a water treatment plant has been developed. The benefits of the system include:

- Only a single calibration is required (focusing of the camera).
- Camera and other instruments are all commercially available, which ensures low material costs.
- The system is compact and easy to install.
- Since the enclosure is pressurized and sealed, the system is applicable to dusty, dirty, and humid industrial environments, as long as the window through which the images are captured is kept clean.

The main conclusions from testing the industrial camera system in the flocculation unit and analyzing the data are as follows:

- Image analysis enables the monitoring of different floc properties and therefore indirect estimation of floc quality.
- The system enables both online and long-term monitoring, because it provides online information on the process, and trend lines can be used to monitor changes occurring during a longer period of time.
- The system can be programmed to alarm in case there are unwanted trends in any of the quality parameters. A warning signal could be delivered to process operators, so that they could check the condition of the floc by naked eye.
- Based on the preliminary results of data analysis, it seems that there are dependencies between the surface area of the floc and certain process measurements, which suggests that it is possible to create data-based models for floc quality.
- Quality models can reveal interesting factors affecting flocculation and help in studying physical phenomena behind the complex process.

**Funding**
The writing of this paper was supported by Maa- ja Vesiteknikan tuki ry. The material on which it is based was produced in the POLARIS project financed by the Finnish Funding Agency for Technology and Innovation (Tekes), which the authors thank for its financial support. In addition, Mika Liukkonen is grateful to the Finnish Cultural Foundation for financial support.

**Author details**
Petri Juntunen
E-mail: petri.juntunen@kuopionvesi.fi
Mika Liukkonen
E-mail: mika.liukkonen@uef.fi
Markku Lehtola
E-mail: markku.lehtola@kuopionvesi.fi
Yrjö Hiltunen
E-mail: yrjo.hiltunen@uef.fi

1 Department of Environmental Science, University of Eastern Finland, Kuopio FI-70210, Finland.

**Citation information**
Cite this article as: Characterization of alum floc in water treatment by image analysis and modeling.

**References**


© 2014 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 3.0 license.