British Journal of Educational Technology Vol 50 No 6 2019 3391–3404
doi:10.1111/bjet.12741

Using automatic image processing to analyze visual artifacts created by students in scientific argumentation

Bo Pei, Wanli Xing and Hee-Sun Lee

Bo Pei is a PhD student in Educational Psychology and Technology. He earned his bachelor's and master's degrees in computer science. His research interests include educational data mining, image processing and machine learning. Wanli Xing is an assistant professor in Instructional Technology at Texas Tech University, USA with background in learning sciences, statistics, computer science and mathematical modeling. His research interests are educational data mining, learning analytics and CSCL. Hee-Sun Lee is a senior research scientist in the Concord Consortium. Her main research areas include technology-enhanced, inquiry-based science curriculum design and evaluation, construct modeling, instrument design, science assessment validation and learning analytics. Address for correspondence: Wanli Xing, Department of Educational Psychology & Leadership, Texas Tech University, Lubbock, TX 79409, USA. Email: wanli.xing@ttu.edu

Abstract

Science classes should support students' development of scientific argumentation. While previous studies have analyzed argumentative texts, they have overlooked the ways in which other types of representations, including images, affect the production of such texts. In addition, studies into the use of visual images in science education have offered mostly qualitative analyses. To fill these gaps in the research, this study used techniques of automated image processing to extract relevant information from student-generated visual artifacts. Specifically, it used a series of image-processing algorithms to automatically extract and quantify features of images created by students to serve as evidence in support of scientific arguments. Using various statistical analyses, we identified the relationships between the extracted features and the students' performance levels in constructing scientific arguments. The results revealed that the presence of water in a student's image correlated significantly with that student's claim and explanation scores and that the amount of water present in a student's image correlated significantly with that student's claim score, but not with their explanation score. These results indicate that automatic image processing can successfully identify image features that affect students' performance in scientific argumentation. Using this analysis as an example, we discuss implications for incorporating automated image processing into further research into scientific argumentation and the development of automated feedback.

Introduction

Scientific argumentation is an essential practice through which scientific knowledge is generated, communicated and refined (Schauble, 2018). By engaging in scientific argumentation, scientists generate evidence-based claims with theoretical backing, examine the constraints of their investigations and critique their own and others' ideas (Nowell, Norris, White & Moules, 2017). For students, scientific argumentation provides genuine opportunities to express and

Practitioner Notes

What is already known about this topic

- Scientific argumentation is an important practice wherein data are critically interpreted in light of scientific knowledge.
- Previous studies have focused on analyzing argumentative texts and have overlooked other types of representation, including images.
- Most studies in science education have employed qualitative methods to study the role of images in science learning.

What this paper adds

- Introduce and demonstrate the workflow of automatic image processing.
- Automatic image processing can efficiently identify the relevant features in the images.
- Visual images and artifacts with different features influence students' claims and explanation differently.

Implications for practice and/or policy

- Educators are recommended to analyze and assess students' produced visual artifacts and images to help students' learning of scientific argumentation.
- Automatic image processing techniques can be implemented to assist teachers to understand and analyze the visual images to provide feedback.

develop their understandings of topics in science by generating and selecting evidence, elaborating scientific reasoning and persuading their peers through rhetorical and dialogical discourse (Schauble, 2018). For this reason, studying students' scientific argumentation can shed light on their use in reasoning of data and models associated with science.

Even though scientific argumentation is conducted via written and spoken language, visual representations can improve the clarity and communicability of evidence-based arguments (Spiegelhalter, Pearson, & Short, 2011). However, the role of visual representations in scientific argumentation remains poorly understood (Kerkhoven, Russo, Land-Zandstra, Saxena & Rodenburg, 2016), mainly because most research has examined students' conceptual understandings of images instead of their use of images as tools in epistemic practices like scientific argumentation (Evagorou & Erduran, 2015). Moreover, most studies that have investigated the use of images in science learning have manually and qualitatively characterized small samples of images (eg, Spiegelhalter, Pearson, & Short, 2011). Given the current trend of learning at scale accompanied by technological advancements (eg, electronic drawings, digital photos and three-dimensional models), students can produce plenty of images in every class. As a result, it has become difficult, if not impossible, for teachers to qualitatively examine every image produced by their students and to provide timely feedback.

In response to these challenges, we use automated image processing techniques that are actively advancing in the field of computational science. Automatic image processing has been widely used in science, engineering and medical fields (Miotto *et al.*, 2017), and in education, image processing techniques can automatically and instantly analyze large-scale images. As a result, they can be used to recognize patterns computationally, to investigate the roles of images in student learning and, eventually, to support the design of real-time feedback for students. This paper explores one use of automatic image processing. Because the main purpose of this study was to introduce automatic image processing, we used a scientific argumentation task as a case study (Yin, 1994). This study sought to answer the following research questions (RQs):

RQ1: How can automatic image processing aid in the analysis of student-generated visual artifacts?

RQ2: How can the results of automatic image processing connect to student learning of scientific argumentation?

In responding to these two research questions, we demonstrated how image processing techniques can be used to automatically extract features from student-generated images. We then investigated the role that visual artifacts generated by students play in their scientific argumentation by examining how features extracted from such artifacts influenced the claims and explanations made by a group of students. As expected, our case study demonstrated the effectiveness of image processing and improved our understanding of how visual artifacts generated by students influence their scientific argumentation. As a result, this work lays the foundation for automatically providing feedback to students when they use images as evidence to support their scientific argumentation.

Background

Scientific argumentation is a language-based activity (Hand, Norton-Meier, Gunel & Akkus, 2016) that is conducted both rhetorically (Chen, Park & Hand, 2016) and dialogically (Clark, Sampson, Weinberger, & Erkens, 2007). Numerous studies have analyzed argumentative texts, and many have developed qualitative-coding schemes to examine the nature of argumentative discourse (Ballard, Dixon, & Harris, 2017; Hill & Sharma, 2015) or to identify patterns in written arguments (Henderson *et al.*, 2018; Wu, Wang & Cheng, 2017; Yim & Warschauer, 2017). To the best of our knowledge, research has yet to examine how student-generated images affect students' mastery of scientific argumentation.

In fact, studies examining student-generated images are relatively rare. An ERIC data search revealed that only 2% of the papers in the education literature had an image component in the analysis (Zafar *et al.*, 2015). Moreover, these studies employed only qualitative analyses. For instance, Miyake *et al.* (2010) studied gender stereotypes in science curriculum resources. They identified 2416 resources, but because these resources were too many to analyze, they qualitatively coded a random sample of 327 resources. Similarly, Ketelhut, Dede, Clarke, Nelson, and Bowman (2017) manually analyzed 91 student-generated images, finding that 19% of the seventh-grade students in their sample developed the scientific understanding intended. Because visualization technologies can provide students using digitally platforms (eg, Massive Online Open Courses) numerous opportunities to generate and utilize images, research and instruction will inevitably make use of massive quantities of learner-generated images. Qualitative content analysis may not be efficient enough for this large-scale image situation and manual coding is too slow for students to be provided prompt feedback (Tang, Xing, & Pei, 2018a).

In line with the recent incorporation of big data analytics by many disciplines, the education community is beginning to experiment with computationally intensive methods for conducting learning analytics and assessment (Tang, Xing, & Pei, 2018b). While traditional learning analytics relies on numerical and textual data (eg, Amigud, Arnedo-Moreno, Daradoumis & Guerrero-Roldan, 2017; Tonidandel, King & Cortina, 2018), a subset of learning analytics called "multimodal learning analytics" utilizes diverse sources of data, including data obtained from images, cameras, wearable sensors, biosensors, and eye-tracking and gesture-sensing technologies (Ochoa, Worsley, Weibel & Oviatt, 2016). When purposefully and meaningfully executed, multimodal learning analytics can yield valuable insights into student learning beyond those provided by numerical and textual analyses alone. For this reason, the image processing techniques introduced in this paper will not only contribute to research into scientific argumentation but also benefit multimodal learning analytics as a whole.

The automatic image processing techniques described in this paper were drawn from a broad range of applications in other disciplines. "Image analysis and processing" refers to the use of computational algorithms to enhance digital images and to analyze such images to obtain quantitative data (Gillies, Kinahan, & Hricak, 2015). Image-processing technologies have been widely used in medical research, science and engineering (Cichocki *et al.*, 2015; Gonzalez, Woods, & Eddins, 2004). For example, Müller, Michoux, Bandon, and Geissbuhler (2004) reviewed the use of content-based image retrieval in medical applications in an effort to describe its clinical benefits. Geneletti and Gorte (2003) proposed a hybrid method that combines GIS analysis with the processing of satellite images to detect differences in land use. Li and Ren (2012) proposed a real-time image processing technique that could detect rail defects. In this paper, we will explore techniques for analyzing and processing images in the context of secondary school students' scientific argumentation practice.

Methods

Research context

The corpus of image data analyzed in this study was obtained from an online module named "Will there be enough fresh water?" (http://authoring.concord.org/sequences/98). Hereafter, we will call this module "the water module." The water module was developed and distributed as part of the High Adventure Science (HAS) curriculum project (https://concord.org/our-work/research-projects/high-adventure-science/), which employed five design principles in developing six online curriculum modules:

- Principle 1. Use open-ended, authentic, frontier science topics to frame the modules.
- Principle 2. Acquaint students with working scientists, their research and their use of computer models.
- Principle 3. Use model-based experimentation as the primary means by which students will acquire content.
- Principle 4. Engage students in building simplified dynamic systems.
- Principle 5. Encourage scientific reasoning and argumentation.

The time frame of implementation suggested for the water module was five to six 45-minute class periods. The suggested target population was students between grades 8 and 12. The water module introduced topics related to freshwater availability and sustainability. The water module contained eight scientific argumentation tasks: four involved computer simulation models and four involved scientific data collected by professional scientists.

In this paper, we focused the images and argumentation responses that students generated in completing the water module's first scientific argumentation task. The reason is that the argumentation tasks that involved scientists' data did not involve images. The other three model-based argumentation tasks had structures similar to that of the first argumentation task (described below). Although the salient features of the images differed between these tasks, the general image processing method outlined in the following section could be expected to work for images generated in the other model-based argumentation tasks.

The first scientific argumentation task prompted students to use an interactive groundwater simulation model to investigate how water droplets precipitated and moved across different sediment layers with different porosity and permeability levels. The students were then asked to capture an image from the model, draw on this captured image the longest path a water droplet could take and then submit the image. The students were then asked to write scientific arguments in which they responded to a two-tiered prompt. This prompt asked them to: (1) make a claim in a multiple-choice format and (2) explain their claim in an open-ended response (see Figure 1 for the entire sequence of the scientific argumentation task).

These images and responses analyzed in this study were produced during the previous year's use of the water module. In that year, one middle and six high school teachers from seven states (KY, IN, MN, MT, NC, OH and PA) in the United States participated in the study. Two schools were located in urban settings, two schools were located in suburban settings and three schools were located in rural settings. The work of the students of these teachers ($n = 290$), including the images and responses they generated in completing the water module, was automatically logged into the curriculum server. We then selected data obtained from students who had produced both images and text responses during the first scientific argumentation task (for a total of 183 students in grades 8–12). It was necessary for each student to have produced both images and text responses because our goal was to assess the relationships between features of students' images and the scores they earned on their argumentation responses.

Scoring of the written argument responses

The domain experts who developed the water module and the argumentation task also scored the students' responses. In the "claim" part, correct claims were given a score of "1" and incorrect claims were given a score of "0." The correct claim was "no," which indicated that the groundwater was not trapped. The students' responses to the explanation prompt followed the claim–evidence reasoning framework (Author, withhold for review) and were scored from 0 to 5 as below. Two human experts scored the entire set of explanation responses. The quadratic kappa value was 0.92.

- Score 0: blank or off-task responses
- Score 1: scientifically invalid statements
- Score 2: variations of the claims that did not include data or reasoning
- Score 3: statements that relied on some relevant data or partially developed reasoning
- Score 4: elaborations of data relevant to the claim
- Score 5: elaborations of relevant scientific reasoning

Data included:

- (model) Water droplets fall and soak into the ground.
- (model) Water droplets move through the ground.
- (model) Water droplets cannot penetrate the black layer.
- (model) Water droplets collect above the black layer.

Reasoning included:

- Water cycles through the Earth's system, sometimes slowly.
- The water table can move up, bringing water to the surface.
- Water moves through the ground because of the porosity and permeability levels of sediments.

Image processing for RQ1

As Figure 2 shows, we used image processing to automatically determine the amount of water present in the images. The amount of water was important because while the simulation model initially contained no groundwater, water was gradually added to the ground in the form of rain

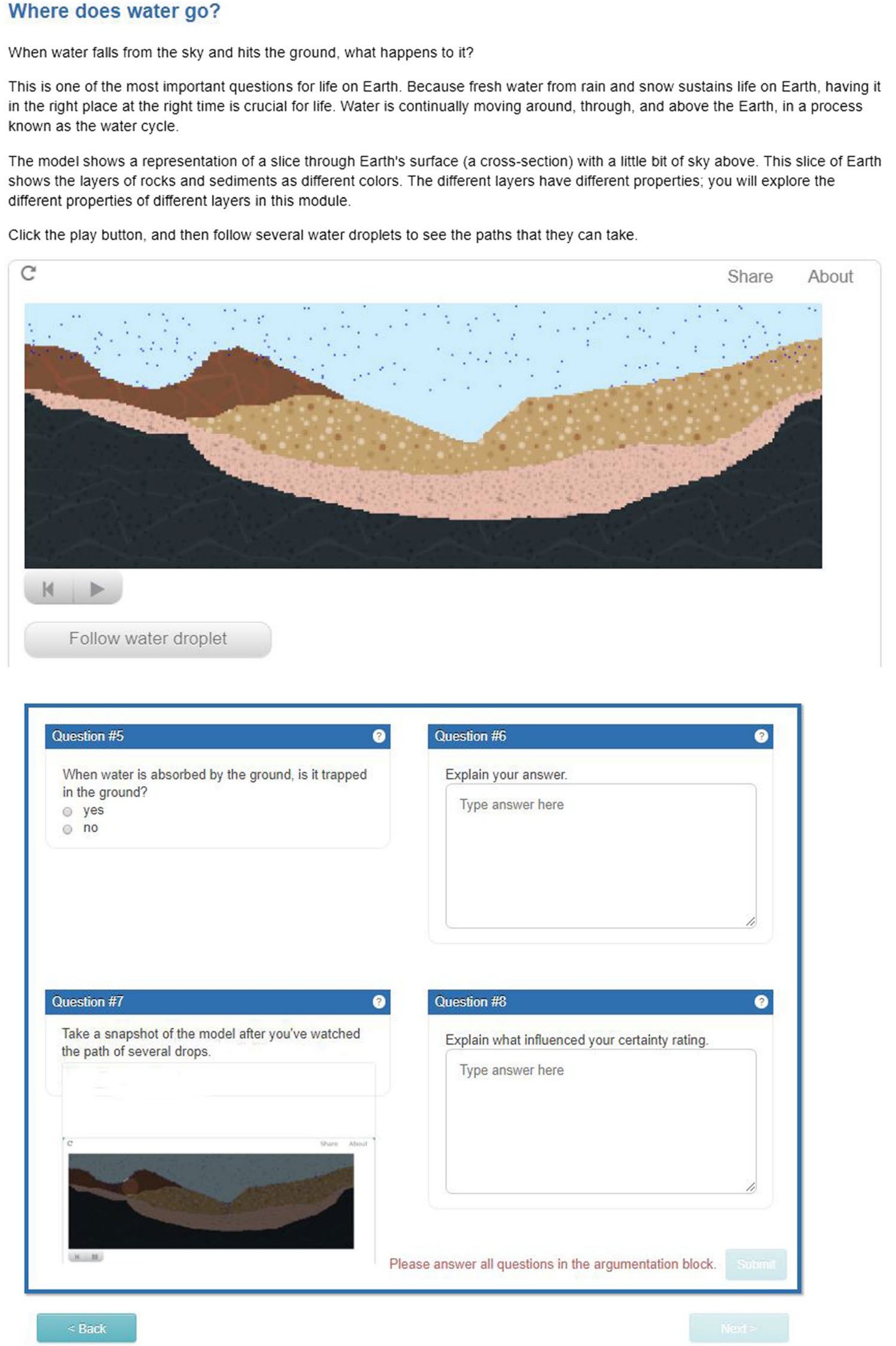

Figure 1: Scientific argumentation task using the groundwater simulation model [Colour figure can be viewed at wileyonlinelibrary.com]

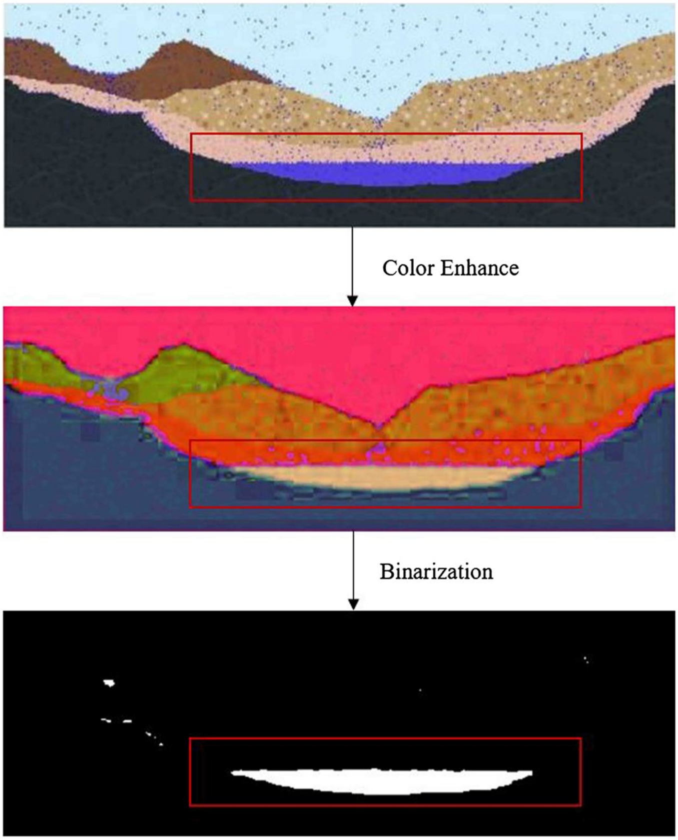

Figure 2: The water extraction process: using a series of image processing techniques to capture the water area (the blue region) in the original image
[Colour figure can be viewed at wileyonlinelibrary.com]

(precipitated water). The rain absorbed into the ground and became groundwater. The amount of water that accumulated in the ground may have affected the students' analyses and understanding of whether the water was trapped underground. One way that the groundwater could escape was through evaporation, which only occurred when the groundwater accumulated to a certain level. As a result, students who watched the model for only a short time may not have observed the evaporation and may have claimed incorrectly that the groundwater was trapped. This response would have constituted an inadequate explanation. Without processing the image, it was difficult to accurately quantify how much water had accumulated as a result of the topography.

The following are the steps we implemented to extract the amount of water in students' images. Although the goal was to quantify the water in pixels in our case, similar steps can be taken to isolate and extract information about any project in images. More detailed information on image processing can be found in Sonka, Hlavac, and Boyle (2014).

1 Image preprocessing ("Color Enhance"): In this step, the goal was to use the region-based brightness enhancement filter to adjust the gray scales of different regions to achieve greater contrast between the regions and thus to distinguish the water region (see the second image in Figure 2). Let $f(i,j)$ be the pixels at position (i,j) in the original image and $g(i,j)$ be the

pixel at position (*i,j*) in the output image. $g(i, j) = \alpha f(i, j) + \beta$ can then be applied to adjust the pixel values in the original image to *g(i,j)*, where $\alpha > 0$ and i and j represent the pixel at (*i,j*). If every pixel in the original image is adjusted, another image is obtained with pixel values different from those of the original and with areas that can be more easily distinguished.

2 Region-of-Interest (ROI) extraction: Once an image with greater interregional contrast was obtained, the region-growing method was applied to automatically separate the water region from the rest. This method allows the extraction of specific regions on the basis of their grayscale values. The first step is to set a threshold value and a seed pixel, ie, the initial pixel point in the region. The second step is to determine the gray values of the pixels neighboring the seed pixel. If the gray values of the neighboring pixels approximate the threshold value (from below or above), the same color is assigned to the two pixel regions. If not, a different color is assigned.

3 Binarization: This process was used to extract the water region. Once a seed was set in the ROI region (ie, the water region) and the color assignment procedure described in step 2 was repeated, the water region was automatically segmented from the neighboring regions. After getting the different regions in the image, a threshold value can be set and the pixel values can be coded either "0"—below the threshold (eg, the black regions in the third image in Figure 2)—or "1"—above the threshold (eg, the white region in the third image in Figure 2). In the last image in Figure 2, the water region is extracted.

4 Feature extraction: In this step, the number of white pixels in the extracted water region (the white region in the third image in Figure 2) was counted. The number of white pixels represented the area of the water region, revealing the amount of water in the image. The black pixels represented the other regions.

5 Evaluation of image processing performance: In this step, a region-based method (Niu *et al.*, 2017) was employed to evaluate the performance of the automatic image processing. Differences were measured between the locations and sizes of the segmented regions. Local Consistency Error (LCE) (Kerkhoven *et al.*, 2016) was used to quantify the degree to which image segmentations were consistent with the ground truth. LCE can be represented as:

$$\mathrm{LCE}\left(\mathrm{S}, \mathrm{R}, p_i\right) = \frac{1}{N} \sum_i \min\left\{E\left(\mathrm{S}, \mathrm{R}, p_i\right), O\left(R, \mathrm{S}, p_i\right)\right\}$$

Table 1: Statistical analyses to determine the relationship between image feature and the students' scientific argumentation scores

	Water vs. no water	*Amount of water*
Claim score	Chi-square test between claim (correct vs. incorrect) and water presence (present vs. not present) using balanced sample size	Independent samples *t* test of the amount of water between incorrect and correct claims
Explanation score	Independent samples *t* test of explanation scores between with water and without water using balanced sample size	ANOVA of the amount of water across truncated explanation scores

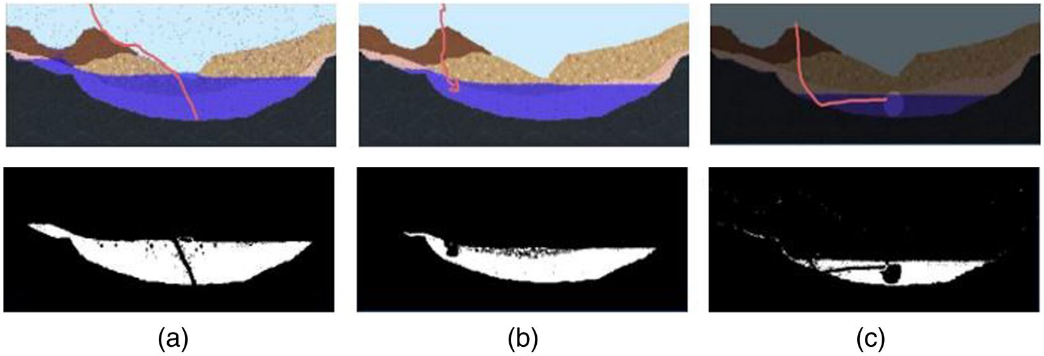

Figure 3: The original images (top) and their corresponding segmentation results (bottom) [Colour figure can be viewed at wileyonlinelibrary.com]

where $E\left(S, R, p\right)$ measures the degree to which the segmentation is consistent with the original image at pixel p, and N is the size of the region to which pixel p belongs. The results of this process revealed the extent to which the automatically extracted water area matched the actual water area.

Statistical analysis for RQ2

Three main methods of statistical analysis were employed to determine the relationship between how much water was present in a student's image and that student's claim and explanation scores as shown in Table 1. We first determined whether students whose images included water earned different claim and explanation scores than did students whose images did not include water. Since the claim scores were binary, we used Chi-square tests to determine whether students whose images included water earned significantly different claim scores than did students whose images did not include water. Since the explanations were scored using a 0–5 scale, independent samples t tests were used to determine whether students whose images included water earned different explanation scores than did students whose images did not include water.

Next, for each student whose image included water, we examined whether the amount of water in their image affected their claim and explanation scores. Independent samples t tests were used to determine whether the amount of water in a student's image significantly affected their likelihood of making a correct claim. In evaluating the relationship between the amount of water in a student's image and that student's explanation scores, we combined the "0," "1" and "2" scores into a single category because the total number of these scores was small. Combining them also made sense because these three explanation scores collectively indicated explanations that contained no data, no reasoning or both. An ANOVA test was then applied to statistically compare with the explanation scores in these four groups. The use of ANOVA was appropriate because of the number of explanation scores (four) being compared to identify significant differences in mean values.

Results

The automatic extraction of image features for RQ1

Figure 3 shows the actual water regions (top) and the automatically segmented water regions (bottom) in different conditions. Note that even though images a–c include different water regions with different noise, the automatically segmented water regions are clearly distinguished in each image. LCE (Local Consistency Error) was further used to quantitatively evaluate the performance of the image processing method. In LCE, a score of "0" indicates that there was no error and a score of "1" represents the maximum error. The average LCE for the images was below .03, indicating that the method of automatic image processing was excellently performed.

Table 2 shows the means and standard deviations of the amounts of water in the students' images after processing and their claim and explanation scores. The results revealed that 41 of the images included no water and 142 included water. In the images that included water, an average of 1470 pixels represented water. The standard deviation was 1708, ranging from 11 pixels to 10 227 pixels.

Presence and amount of water for RQ2

Given that the number of images that included water (n = 142) was much greater than the number of images that did not include water (n = 41), we randomly sampled 41 of the images that included water to balance the sizes of the samples. Table 3 shows the results of the cross-tabulation of "claim score" and "presence of water in the image." As Table 3 shows, we found that the ratio of correct to incorrect claims for the students whose images included water (21/20) was noticeably different than the ratio of correct to incorrect claims for the students whose images did not include water (9/43). This difference was statistically significant: $\chi^2(df = 1) = 4.76$, $p < 0.05$. This finding indicates that the students who cropped the image while playing the simulation model were more likely to have made correct scientific claims than were the students who cropped the image when the model was not playing.

Table 3 also shows the descriptive statistics for the explanation scores. The results revealed that the explanation scores of the students whose images included water ($M = 4.12$, $SD = 1.26$) were higher than the explanation scores of the students whose images did not include water ($M = 3.05$, $SD = 1.8$). This difference was also statistically significant: $t(df = 81) = -3.09$, $p < 0.01$).

For each student whose image included water, we examined the relationship between the amount of water their image contained and their claim and explanation scores. Since one of the students did not complete their scientific argument, we removed them from the analysis. In total, we analyzed 141 images. We found that the mean amount of water in the images generated by the students who made incorrect claims (M = 1795 pixels, SD = 2236 pixels) was higher than the mean amount of water in the images generated by the students who made correct claims (M = 1115 pixels, SD = 834 pixels). The follow-up t test showed a statistically significant difference in these amounts: $t(df = 140) = 2.38$, $p < 0.05$).

Table 4 shows the descriptive statistics for the amount of water for each explanation-score category. The results suggest that the students whose images contained the largest mean amount of water earned the lowest explanation scores ("0," "1" and "2"). In contrast, the students whose images contained the lowest mean amount of water earned the highest explanation scores ("5"). The students whose images contained the second largest mean amount of water earned an explanation score of "4," which indicated the elaboration of relevant data. Follow-up ANOVA tests

Table 2: Descriptive statistics for water amount, claim score and explanation score (n = 183)

	Mean	*Median*	SD	*Min*	*Max*
Water amount	1141	783	1624	0	10 227
Claim	0.50	0.50	0.50	0 (incorrect)	1 (correct)
Explanation	3.44	4.00	1.63	0	5

Table 3: Effect of the presence of water on claim and explanation scores

	Water		*No water*	
Claim distribution	Correct	Incorrect	Correct	Incorrect
	21	20	9	32
Explanation score	Mean	SD	Mean	SD
	4.12	1.26	3.05	1.8

Table 4: Descriptive statistics of water amount in different explanations

	n	*Mean*	*Median*	SD	*Min*	*Max*
Explanation_0, 1, 2	27	1855	1227	2056	31	9046
Explanation_3	25	1292	1010	1712	44	8840
Explanation_4	39	1589	982	1921	84	10 227
Explanation_5	50	1267	917	1305	11	7700

revealed no significant differences in "mean amount of water" among the four explanation-score categories: $F(df = 139) = 1.34$, $p = 0.25$). Since the standard deviation was prone to extremes, we also considered the median amounts of water for the four explanation-score categories. This revealed that the median amount of water decreased as the explanation score increased.

In sum, whether a student's image included water significantly affected both their claim score and their explanation score. While the mean amounts of water for the four explanation-score categories were not significantly different, the median amount of water decreased as the explanation score increased, indicating that students who struggled with explanation tended to run the simulation model longer.

Discussion

In this study, we used automatic image processing to extract features from student-generated images and then determined whether these features correlated with the students' claim and explanation scores on a scientific argumentation task. The results indicated that the presence of water in a student's image correlated significantly with that student's claim and explanation scores. Indeed, students who devoted sufficient time to observing the flow of groundwater earned higher claim and explanation scores than did students who did not. However, students who spent too much time running the simulation model earned lower claim and explanation scores. In fact, the more water appeared in a student's image, the more likely that student was to make incorrect claims. The mean amount of water present in the images did not correlate significantly with the mean explanation score. Nonetheless, the median amount of water decreased as the median explanation score increased. Students who ran the simulation for too long may have been less confident in their claims and less likely to make correct claims.

Image processing techniques have only recently been used to study scientific argumentation. While some studies have examined the role of visual images in scientific argumentation, these studies have generally used qualitative methods in describing parts of images and drawing conclusions (Bricker & Bell, 2008; Forbes & Davis, 2008). In contrast, this study took advantage of recent advancements in image processing technologies (Sonka, Hlavac, & Boyle, 2014) to automatically process images that students created as evidence for their claims in a scientific argumentation task. Specifically, we constructed a workflow that accurately isolated and quantified a region of interest in the students' images (the region containing water). This generic image processing workflow could be used by other researchers studying images in science learning. An added benefit of automated image processing is that it enables large numbers of images to be analyzed and scored. An automated image processing engine could be developed and embedded into a digital curriculum module to diagnose images and provide real-time feedback to students. In a setting similar to the one in this study, eg, if an image processing engine detected that there was no water at all in a student's image, an alert message could be automatically sent to that student suggesting that they observe how the water flows underground before they begin writing. If the image processing engine detected that there was too much water in a student's image, however, it could indicate that the student was having difficulty understanding how the water flowed underground.

The learning analytics community rarely considers images in attempting to capture students' thinking. However, when combined with multimodal data sources—eg, videos, texts and clickstreams—the method of image analysis presented in this study could yield new insights into the visual pathway of student learning. Furthermore, this study demonstrated only how automatic feature extraction can be used to study students' learning of argumentation. Other image processing applications, such as image matching, can be used to advance research into student learning (Karami, Prasad, & Shehata, 2017). For example, concept mapping is widely used in education, and image matching could be used to automatically compare concept maps created by students with those created by experts. Such comparisons could help teachers develop and implement intervention strategies.

While this study offers a concrete example of automatic image processing, it does have limitations. Image processing can work very well in bounded educational scenarios, eg, with images that have predefined features or are made according to specific instructions. In unbounded educational scenarios, however, eg, when students are asked to draw freely with no instructions or constraints, then the area of interest in the images may become too complex to define, limiting the ability of automatic image processing technologies to extract relevant features. Another limitation of this study is that its results are correlational. Additional controls and qualitative interviews would be necessary to determine why and how features of images influence students' thought processes and performance outcomes.

Conclusion

In this study, we used image processing technologies to automatically extract features from student-generated images and examined how these features influenced the students' performance levels in scientific argumentation. In contrast to existing manual and qualitative image analyses, the image processing technologies used in this study were able to automatically distinguish the presence and quantity of water in student images. The following statistical analyses were used to determine whether and how these extracted features correlated with student learning outcomes. These findings have practical implications for supporting students' science learning with images. Several directions for future research are noted. First, since this study found that the amount of water in a student's image correlated with their performance in scientific argumentation, an immediate follow-up could be to identify the threshold value for water quantity that most improves student performance. Second, future work could integrate the image processing procedures into the HAS digital curriculum module as an input to the intelligent feedback system (Zhu *et al.*, 2017) and study its effects on student learning. Third, to support the generalizability of this study, additional research could apply similar image processing methodologies to other scientific argumentation tasks involving student-created images.

Statements on open data, ethics and conflict of interest

The data set was collected with IRB approval from the authors' organization. These data will only be made available to other researchers if specific requests for amendments are made to the current approvals and will be considered on a case-by-case basis.

The study was conducted in accordance with BERA Ethical Guidelines. No personal identifiers were reported in this study.

There is no potential conflict of interest in the work.

References

Amigud, A., Arnedo-Moreno, J., Daradoumis, T., & Guerrero-Roldan, A. E. (2017). Using learning analytics for preserving academic integrity. *The International Review of Research in Open and Distributed Learning*, *18*(5), 104–119.

Ballard, H. L., Dixon, C. G., & Harris, E. M. (2017). Youth-focused citizen science: Examining the role of environmental science learning and agency for conservation. *Biological Conservation*, *208*, 65–75.

Bricker, L. A., & Bell, P. (2008). Conceptualizations of argumentation from science studies and the learning sciences and their implications for the practices of science education. *Science Education*, *92*(3), 473–498.

Chen, Y. C., Park, S., & Hand, B. (2016). Examining the use of talk and writing for students' development of scientific conceptual knowledge through constructing and critiquing arguments. *Cognition and Instruction*, *34*(2), 100–147.

Cichocki, A., Mandic, D., De Lathauwer, L., Zhou, G., Zhao, Q., Caiafa, C., & Phan, H. A. (2015). Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine*, *32*(2), 145–163.

Clark, D., Sampson, V., Weinberger, A., & Erkens, G. (2007). Analytic frameworks for assessing dialogic argumentation in online learning environments. *Educational Psychology Review*, *19*, 343–374.

Evagorou, M., & Erduran, S. (2015). The role of visual representations in scientific practices: From conceptual understanding and knowledge generation to 'seeing' how science works'. *International Journal of STEM Education*, *2*(1), 11.

Forbes, C. T., & Davis, E. A. (2008). Exploring preservice elementary teachers' critique and adaptation of science curriculum materials in respect to socio-scientific issues. *Science & Education*, *17*(8–9), 829–854.

Geneletti, D., & Gorte, B. G. H. (2003). A method for object-oriented land cover classification combining Landsat TM data and aerial photographs. *International Journal of Remote Sensing*, *24*(6), 1273–1286.

Gillies, R. J., Kinahan, P. E., & Hricak, H. (2015). Radiomics: Images are more than pictures, they are data. *Radiology*, *278*(2), 563–577.

Gonzalez, R. C., Woods, R. E., & Eddins, S. L. (2004). *Digital image processing using MATLAB* (2rd ed.). New Jersey, Upper Saddle River: Pearson-Prentice-Hall.

Hand, B., Norton-Meier, L. A., Gunel, M., & Akkus, R. (2016). Aligning teaching to learning: A 3-year study examining the embedding of language and argumentation into elementary science classrooms. *International Journal of Science and Mathematics Education*, *14*(5), 847–863.

Henderson, J. B., McNeill, K. L., González-Howard, M., Close, K., & Evans, M. (2018). Key challenges and future directions for educational research on scientific argumentation. *Journal of Research in Science Teaching*, *55*(1), 5–18.

Hill, M., & Sharma, M. D. (2015). Students' representational fluency at university: A cross-sectional measure of how multiple representations are used by physics students using the representational fluency survey. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(6), 1633–1655.

Karami, E., Prasad, S., & Shehata, M. (2017). Image matching using SIFT, SURF, BRIEF and ORB: Performance comparison for distorted images. *arXiv preprint arXiv:1710.02726*.

Kerkhoven, A. H., Russo, P., Land-Zandstra, A. M., Saxena, A., & Rodenburg, F. J. (2016). Gender stereotypes in science education resources: A visual content analysis. *PLoS ONE*, *11*(11), e0165037.

Ketelhut, D. J., Dede, C., Clarke, J., Nelson, B., & Bowman, C. (2017). Studying situated learning in a multiuser virtual environment. In E. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil (Eds.), *Assessment of problem-solving using simulations* (1st ed., pp. 47–68). New York, NY: Routledge.

Li, Q., & Ren, S. (2012). A real-time visual inspection system for discrete surface defects of rail heads. *IEEE Transactions on Instrumentation and Measurement*, *61*(8), 2189–2199.

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, *6*(19), 1236–1246.

Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, *330*(6008), 1234–1237.

Müller, H., Michoux, N., Bandon, D., & Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications—Clinical benefits and future directions. *International Journal of Medical Informatics*, *73*(1), 1–23.

Niu, S., Chen, Q., De Sisternes, L., Ji, Z., Zhou, Z., & Rubin, D. L. (2017). Robust noise region-based active contour model via local similarity factor for image segmentation. *Pattern Recognition*, *61*, 104–119.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, *16*(1), 1609406917733847.

Ochoa, X., Worsley, M., Weibel, N., & Oviatt, S. (2016, April). Multimodal learning analytics data challenges. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 498–499). New York, NY: ACM.

Schauble, L. (2018). In the eye of the beholder: Domain-general and domain-specific reasoning in science. In F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific reasoning and argumentation* (1st ed., pp. 11–33). New York, NY: Routledge.

Sonka, M., Boyle, R., & Hlavac, V. (2016). *Image processing, analysis, and machine vision* (4th ed.). Florence: Cengage Learning.

Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future. *Science*, *333*(6048), 1393–1400.

Tang, H., Xing, W., & Pei, B. (2018a). Time really matters: Understanding the temporal dimension of online learning using educational data mining. *Journal of Educational Computing Research*, 1–22, 0735633118784705.

Tang, H., Xing, W., & Pei, B. (2018b). Exploring the temporal dimension of forum participation in MOOCs. *Distance Education*, 1–20.

Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods*, *21*(3), 525–547.

Wu, Y. T., Wang, L. J., & Cheng, T. Y. (2017). *A preliminary study of university students' collaborative learning behavior patterns in the context of online argumentation learning activities: The role of idea-centered collaborative argumentation instruction*. Philadelphia, PA: International Society of the Learning Sciences.

Yim, S., & Warschauer, M. (2017). Web-based collaborative writing in L2 contexts: Methodological insights from text mining. *Language Learning & Technology*, *21*(1), 146–165.

Yin, R. K. (1994). Discovering the future of the case study. Method in evaluation research. *Evaluation practice*, *15*(3), 283–290.

Zafar, S. N., Shah, A. A., Hashmi, Z. G., Efron, D. T., Haut, E. R., Schneider, E. B., … Haider, A. H. (2015). Outcomes after emergency general surgery at teaching versus nonteaching hospitals. *Journal of Trauma and Acute Care Surgery*, *78*(1), 69–77.

Zhu, M., Lee, H.-S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education*, *39*(12), 1648–1668.