Quality Control of Geological Voxel Models using Experts’ Gaze

Peter-Paul van Maanen*, Freek S. Busschersb, Anne-Marie Brouwera, Michiel J. van der Meulenb, Jan B. F. van Erpa

aNetherlands Organisation of Applied Scientific Research (TNO)
bTNO Geological Survey of the Netherlands

Abstract

Due to an expected increase in geological voxel model data-flow and user demands, the development of improved quality control for such models is crucial. This study explores the potential of a new type of quality control that improves the detection of errors by just using gaze behavior of 12 geological experts. Gaze is used as input for an attention model that results in attended areas on sliced representations of part of a geological voxel model. We compared attended areas to errors as manually marked by the experts. We found a clear match between manually marked errors and attended areas as determined using gaze. We also found that a large proportion of this match is reached within a small amount of total viewing time.

Keywords: Geological Voxel Model, Quality Control, Gaze, Attention Model

Email addresses: peter-paul.vanmaanen@tno.nl (Peter-Paul van Maanen), freek.busschers@tno.nl (Freek S. Busschers), anne-marie.brouwer@tno.nl (Anne-Marie Brouwer), michiel.vandermeulen@tno.nl (Michiel J. van der Meulen), jan.vanerp@tno.nl (Jan B. F. van Erp)
1. Introduction

Geological voxel models are predictions of the architecture and properties of the subsurface in 3D. Especially when undertaken at national scale, geological voxel modeling uses and produces vastly more data and information than traditional 2D geological mapping (van der Meulen et al., 2013). This in its turn presents new challenges, since it surpasses the capabilities of current model quality control.

It is generally accepted that errors will draw experts’ attention and visual attention is closely intertwined with gaze location (Rizzolatti et al., 1987; Corbetta et al., 1998; Carpenter, 1988; Land and Furneaux, 1997) and gaze duration (e.g., Brouwer et al., 2013). We devised an experiment that captures eye gaze behavior of 12 geological experts who were asked to visually check a geological voxel model for errors, and then also manually mark the errors they identified. In this way, we explored the effectiveness and reliability of a novel, potentially faster method to check the model relying on gaze alone.

Other than speed, a potential advantage of using gaze data analysis over conventional error reporting is that it circumvents conscious deliberations of the experts. These can be related to (unconscious) reluctance to manually report errors due to own involvement with the geological model. More importantly, an expert may be reluctant to mark a feature as erroneous if he or she cannot explain why, i.e., if it is only a ‘feeling’ that something is wrong. However, even if observers do not consciously recognize anomalies when gazing at anomalous objects, gaze duration tends to be longer (Droll et al., 2005; Hayhoe et al., 1998). If this holds true, then quality control based on gaze would unlock expert intuition and experience in a new way. To explore the
potential of this new type of quality control, we define the following research questions:

RQ1 To what extent do attended areas as determined from experts gaze data match with geological model errors as subsequently indicated by the same experts using a mouse?

RQ2 How does this match change over time during geological model quality control?

2. Geological Model

The geological voxel model used in this study is GeoTOP (Stafleu et al., 2011). GeoTOP is a schematization of the subsurface using voxels (‘3D pixels’) of 100 by 100 meters in horizontal directions and 0.5 meters in the vertical direction (see Fig. 1). Each voxel has estimates of stratigraphy and lithology: clay, sand (in three grain-size classes), gravel, peat or ‘other’ (see for example Fig. 2).

The GeoTOP modeling procedure consists of automated database queries, 2D modeling of stratigraphic surfaces and 3D property modeling, designed in such a way that it provides the best possible representation of the geological features, given the available data and expert knowledge. The work-flow includes quality checks on both input data and modeled output, and the supply of 2D and 3D uncertainty estimates. However, errors in the geological plausibility of the modeled output are much more difficult to capture. Whether for example variation in position of a bounding surface or the geometry of a fluvial sand body is geologically realistic is very difficult to assess using
Figure 1: Voxel modeling work-flow (van der Meulen et al., 2013). Borehole information is coded lithostratigraphically and as lithological classes (Step 1). 2D modeling of basal lithostratigraphical unit boundaries (Step 2). 3D modeling of lithological classes (Step 3).
Figure 2: Block diagram showing part of the GeoTOP model output for the Utrecht and Gelderse Vallei area (surface area: 62× 24 km; model base: 50 m below Dutch Ordnance Datum; vertical exc. 75×). Colors represent different GeoTOP model lithostratigraphic units. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)
computational algorithms only. Checking models for this type of errors is therefore carried out on a manual basis by geological experts, which is an extremely time consuming process.

3. Attention Model

The model used for estimating attended areas from gaze data is an attention model, based on the dynamical model of visual attention as described by Bosse et al. (2009, 2012). The model produces attention values for each predefined area (pixel) of an examined image. For every point in time during inspection, an attentional unit is divided across pixels of the image where the pixel at the center of gaze receives the largest value while pixels surrounding this pixel receive a value that decreases with the distance they are from the center of gaze. This reflects the idea that while people usually attend to the gaze location rather than to the visual periphery, attention is not directed at a single pixel but at an area. In addition, gaze history is taken into account. The attention model includes a decay function that represents the rate at which attention as given to a single pixel dissipates over time. This ensures that locations that are gazed at for a long time are considered to have received attention while locations that are briefly skipped over are not considered to have received attention. The attention model has three free parameters. Parameter $\gamma$ defines the rate at which attention decreases with increasing distance from the center of gaze. Parameter $\lambda$ defines the rate at which attention decays over time. Finally, parameter $\alpha$ is a threshold used to convert the gradual output of the attentional model into binary values reflecting whether an observer attended a specific area or not.
3.1. Attention Value

The attention model of Bosse et al. (2009, 2012) defines different (discrete) areas over visual stimuli (e.g., images, movies and displays), which each have a specific quantity of attention on each time point. This quantity is called the attention value. The sum over all quantities in each area is assumed to be constant:

\[ A(t) = \sum_{x,y} AV(x, y, t) = 1 \]  (1)

where \( A(t) \) is the total amount of attention at time point \( t \) and \( AV(x, y, t) \) is the attention value for area \((x, y)\) at time point \( t \). Areas are defined as \( 1 \times 1 \) squares within an \( M \times N \) grid spread over each visual stimulus.

3.2. Error

When an expert is checking for errors, we expect that higher attention values indicate suspicious areas. That is, an area \((x, y)\) is estimated as erroneous if the attention value for \((x, y, t)\) exceeds a certain threshold \(\alpha\):

\[ e(x, y) = \begin{cases} 
1 & \text{if } AV(x, y, t) \geq \alpha \text{ for some } t \\
0 & \text{otherwise}
\end{cases} \]  (2)

where \( e(x, y) \) is a binary estimation value indicating whether \((x, y)\) is erroneous. For each area \((x, y)\) we test whether this holds by taking the maximum of \( AV(x, y, t) \) for all time points \( t \) during the presentation of a visual stimulus and verify whether it exceeds \(\alpha\).
3.3. Gaze

Since people pay more attention to the center than to the periphery of their visual space, the relative distance of each area \((x, y)\) to the gaze point (the center) is an important factor in determining the attention value of \((x, y)\). For each new attention value over time, this is modeled as follows:

\[
AV_{\text{new}}(x, y, t) = \frac{1}{1 + \gamma \cdot r(x, y, t)^2}
\]  

(3)

where the term \(r(x, y, t)\) is taken as the Euclidean distance between the gaze point at time point \(t\) and \((x, y)\), multiplied by a factor \(\gamma\) which determines the relative impact of the distance to the gaze point on the attentional state:

\[
r(x, y, t) = \sqrt{(x_g(t) - x)^2 + (y_g(t) - y)^2}
\]  

(4)

where \(x_g(t)\) and \(y_g(t)\) are the coordinates of the gaze point at time point \(t\).

3.4. Normalization

The total amount of human attention per time point is limited. Therefore the attention value for each area \((x, y)\) is expressed as a proportion of the attention values summed over all areas, with the maximum attention value of an area being 1 and the minimum being 0:

\[
AV_{\text{norm}}(x, y, t) = \frac{AV_{\text{new}}(x, y, t)}{\sum_{x', y'} AV_{\text{new}}(x', y', t) \cdot A(t)}
\]

and, since \(A(t) = 1\) from Eq. 1, it holds that:

\[
AV_{\text{norm}}(x, y, t) = \frac{AV_{\text{new}}(x, y, t)}{\sum_{x', y'} AV_{\text{new}}(x', y', t)}
\]  

(5)

where \(AV_{\text{norm}}(x, y, t)\) is the normalized attention value for area \((x, y)\) at time point \(t\).
3.5. Decay

In order to move from one attentional area to the next, gaze travels across the image more quickly than attention does, i.e., attention is less dynamic than gaze (Theeuwes, 1994). In order to capture this, the model includes a decay factor which allows attention values to persist over time:

\[
AV(x, y, t) = \lambda^{t' - t} \cdot AV(x, y, t') + (1 - \lambda^{t' - t}) \cdot AV_{norm}(x, y, t)
\]

where \( \lambda \) is the decay parameter that results in the gradual decay of the attention value of \((x, y)\) at time point \(t'\), and gradual introduction of the normalized attention value of \((x, y)\) at time point \(t\), where \(t' < t\). Note that higher values for \( \lambda \) result in a higher persistence and lower decay and vice versa.

4. Method

4.1. Participants

Twelve expert geologists of the Dutch geological survey (ten male, two female) participated in the experiment on a voluntary basis. The average age was 42.5 ± 10.7 years. Six participants hold a MSc degree and five also a PhD degree in earth sciences or a related field, one received vocational training. All participants had experience in the type of geology related to the used experimental stimuli, although each with a different subarea of expertise. The average experience with using or validating geological models was 7.3 ± 3.8 year.
4.2. Stimuli

For the experiment we used part of the GeoTOP model from an area in the west of the Netherlands. The model block has a dimension of 20 700 m (x) by 24 500 m (y) and provides geological information up to a depth of 50 m below Dutch ordnance datum (z). The geology consists of ~30 m of Pleistocene Rhine-Meuse fluvial sand covered by ~5 m Holocene flood (clay and peat) and channel belt sands of the Rhine and Meuse. In the northern part of the model block, fluvio-glacial sediments and glacio-tectonized sediments (Utrechtse Heuvelrug) occur. The model block is representative for shallow subsurface geological variability and complexity in the central part of the Netherlands. In this experiment we used the lithostratigraphic voxel attribute for testing. Stimuli consisted of 70 images of 980 × 480 pixels each. These images are consecutive (y, z)-slices of the model block (see Fig. 3).

4.3. Materials

Images were presented in the center of a 1024 × 768 pixels monitor and viewed from a distance of 61 cm. To minimize head movements a chin rest was used. Eye movements were recorded using a Tobii x50 Eye Tracker with a sample rate of 50 Hz, accuracy of 0.5 degrees (deviation between the measured and actual gaze point of the user), and spatial resolution of 0.35 degrees (frame-to-frame variation). Long term drift and head movement are compensated for automatically by the eye tracker.

4.4. Task

The participants were asked to look for errors in a part of GeoTOP. Following common practice when checking geological voxel models, they scanned
Figure 3: GeoTOP model lithostratigraphic units in an area in the west-central Netherlands: a) the entire model block, b) an example of a single slice, c) the heat map showing the distribution of the areas indicated as ‘erroneous’ by the participants and d) a post hoc annotation of these areas. The participants’ indication of errors in the different post hoc annotations are explained as: 1) pointed base of gray unit, 2) irregularity in base of light gray unit, 3) sharp lateral transition between light brown and gray unit, irregular base of orange unit, 4) pointed base of orange unit, 5) irregular top of gray unit, 6) sharp lateral transition between yellow and light red unit, 7) absence of light gray unit resulting in high occurrence of light red unit, irregularities in base of orange unit, 8) sharp lateral transition between dark red and orange unit, pointed base of dark yellow unit, deep occurrence of light blue unit, 9) light yellow unit positioned on top of dark blue unit, 10) light yellow unit positioned on top of dark blue unit, high occurrence of light blue unit, 11) light yellow unit positioned on top of dark blue unit, shape of dark blue unit and 12) shapes of dark blue unit. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)
through consecutive slices of this part of GeoTOP using the keyboard. The participants indicated when they were done reviewing a slice by pressing the space bar, therewith finishing a ‘view interval’. After that, the same slice was presented again and they marked presumed erroneous areas by painting them black, using the mouse. When they were satisfied with their markings, they again pressed the space bar, therewith finishing a ‘mark interval’. After this, the next slice appeared, until all slices were finished. The complete task took on average 70 minutes to complete and there was no break. Each view interval took on average 29 seconds and each mark interval on average 31 seconds.

4.5. Procedure

Before the task started, participants filled out a form stating their age and years of experience with geological data, mapping and model validation. They also read and signed an informed consent form. The chin rest and chair were adjusted such that the participant could sit as comfortably as possible throughout the experiment. The eye tracker was calibrated using a nine-point calibration (using the software “Clearview”).

The participants then went through an example slice in order to become familiar with the task interface. Instructions stressed that participants should only search for erroneous areas during the view interval and that during the mark interval only those areas should be marked. In order to prevent problems with the interpretation of the marked errors, instructions also stressed that participants should not only circle around erroneous areas, but completely paint them black. Any questions about the presented slice and interface were answered. Then the experiment started.
After the experiment, participants filled out a questionnaire related to the perceived difficulty of the task, their expected performance as determined from the ‘mark intervals’ and experienced mental workload (on ten-point scales). They were also asked to estimate after how much time (seconds) they thought in general they had determined 80% of the indicated errors in the view intervals.

4.6. Analysis

4.6.1. Match between Modeled and Marked Errors (RQ1)

For each slice, the attention model as described in Section 3 resulted in an attentional map fluctuating over time. Using Eq. 2, the geological model errors as estimated by the attention model were calculated for each slice. These were compared with the participants’ markings of erroneous areas, as made during the subsequent mark interval. Fig. 4 shows an example of the erroneous areas as estimated by the attention model (black) and as manually indicated (white). It shows the match between the two for a given slice, a given geological expert and given settings of the three free attention model parameters.

As the model has three free parameters ($\alpha$, $\gamma$ and $\lambda$, for attentional threshold, decay of attention with distance from the center of gaze and decay of attention over time, respectively), the above-mentioned match was optimized using linear search parameter tuning. For this purpose a ‘good match’ needs to be defined. Parameters can be set to accommodate a large hit rate, i.e. such that the attention model misses the least amount of pixels on each slice representing geological model errors as indicated manually. However, this will inevitably lead to a relatively high number of false alarms, i.e., a
Figure 4: Attentional map at the end of the view interval for the slice as seen in Fig. 3b for an arbitrary participant and attention model parameters with the resulting modeled erroneous areas in black (a) and the participant’s markings of erroneous areas at the end of the mark interval in white (b).
large number of geological model errors as indicated by the attention model that are not marked as erroneous areas. Setting parameters to minimize the number of false alarms will increase the number of misses. The relative importance of maximizing hits and minimizing false alarms will depend on the application purpose. Therefore, we optimized the three free parameters for a range of different hit rates (from 0 to 1 in 101 equidistant steps), each time searching for settings that minimized the associated false alarm rate. The optimization procedure followed was leave-one-out cross-validation over all 70 slices for each participant and for 101 different hit rate settings. This implies that for each participant, parameter settings were optimized using data from 69 slices. The attention model was then applied to the remaining (evaluation) slice and hits and false alarms were determined for that particular slice. This was repeated until all slices were used as evaluation slice (i.e., repeated 70 times). The optimization of the model parameter $\alpha$ was done with a granularity of 101 and that of $\gamma$ and $\lambda$ with a granularity of 31. This means that $101 \times 31 \times 31 = 97061$ different model parameter settings were evaluated per optimization procedure.

This procedure resulted in a so-called receiver operating characteristic curve (ROC curve), showing the relationship between hit rates and false alarm rates, with $101$ (hit rate settings) $\times 12$ (participants) $= 1212$ (false alarm rate, hit rate)-pairs through which a polynomial curve was fit. The point furthest away from the line of no-discrimination (line between $(0,0)$ and $(1,1)$) represents the best model performance (sensitivity index $d'$) and model parameters, for the special case that the valuation of higher hit rates is the same as for lower false alarm rates. To determine overall model per-
formance, independent of higher hit versus lower false alarm rate valuation, we calculated the area under the ROC-curve (AUC):

$$\text{AUC}_{RQ1} = \int_{0}^{1} h(f) \, df \quad (7)$$

where $h(f)$ is the hit rate for a particular false alarm rate $f$. AUCs larger than 0.5 (ROC-curve above the line of no-discrimination) show a match between geological model errors as indicated by the attention model and by manual marking that is larger than chance (answer to RQ1).

4.6.2. Match for Increasing View Interval Durations (RQ2)

To answer the second research question, the match between the geological model errors as indicated by the attention model and manual markings was evaluated for increasing time cut-off points in the view intervals (i.e., increasing view interval durations). Thus the gaze data for each slice for each participant was limited to an increasing amount of time, as opposed to the complete view time as used for RQ1.

The AUC for each participant (as described for RQ1) was calculated using 146 different view interval durations, ranging from 0 seconds to 29 seconds, where the latter is the average time the participants spend during view intervals. Parameter tuning was done for each view interval duration, because we expected differences in the dynamics of attention over time, which would also result in different optimal parameter settings. For relative comparisons between different view interval durations, the AUCs were compensated for performances reached by chance. This was done by subtracting 0.5 (the expected AUC for random models) from each AUC. The AUCs per view interval duration for RQ2 are therefore calculated in the following way:
AUC_{RQ2}(t) = \int_0^1 h(t)(f(t)) \, df(t) - 0.5 \quad (8)

where $t$ is the view interval duration ranging from 0 to 29 seconds in steps of 0.2 seconds (146 steps in total).

Based on the above AUC_{RQ2}-calculations, a plot with 146 (view interval durations) $\times$ 12 (participants) = 1752 (view interval duration, AUC_{RQ2})-pairs was constructed, through which a polynomial curve was fitted and compared to the case that the model match improves linearly over time. This linearly improving match is represented by a line plotted from no gaze data and random model match (i.e., coordinate (0, 0)) to all gaze data and the model match as obtained after a view interval duration of 29 seconds (i.e., coordinate (29, AUC_{RQ2}(29))). If the fitted curve is above this line, specifically for shorter view interval durations, the attention model matched the manual markings of geological model errors in a fraction of the time geologists used for detecting errors (answer to RQ2).

To check whether the above procedure for tuning parameters per view interval duration is necessary (apart from tuning per participant and hit rate setting), the dependency of the model parameter settings for $\gamma$, $\lambda$ and $\alpha$ is shown by plotting their respective values for each view interval duration. Systematic fluctuations of these values show this dependency. A comparison between the AUC_{RQ2}s with and without tuning per view interval duration shows the need for tuning per view interval duration with respect to model match. This comparison was done by calculating the AUC of fitted curves on the AUC_{RQ2}s with and without tuning per view interval duration, where the model parameters without tuning per view interval duration were set to the
optimal model parameters when tuning based on the gaze data of the complete view interval durations per participant per hit rate setting. This means that $12 \times 101 = 1212$ model parameter settings were used without tuning per view interval duration and $12 \times 101 \times 146 = 176952$ model parameter settings with tuning per view interval duration.

5. Results

5.1. Questionnaire

Participants’ perceived difficulty of the task was 3.7$\pm$1.5 (i.e., fairly easy), the expected performance was 6.8 $\pm$ 1.7 (fairly good) and the experienced mental workload was 7.4$\pm$1.0, which can be considered as high. Participants varied widely in their estimates as to how much time they needed to find 80\% of the errors: responses ranged from 5 to 50 seconds with an average of 15.6 $\pm$ 12.9 seconds (i.e., 53.8 $\pm$ 44.5\% of 29 seconds).

5.2. Match between Modeled and Marked Errors (RQ1)

Fig. 5 shows the ROC curve for different optimal parameter settings of the model for varying hit rate settings. Participants can be discerned in the dotted ROC curves. The blue curve depicts the average. The green cross indicates the point that is furthest away from the line of no-discrimination, with a false alarm rate of 0.18, a hit rate of 0.68, and a $d' = 0.68 - 0.18 = 0.50$. This means that 68\% of the area manually marked as erroneous was detected using the attention model, and 18\% of the area not manually marked as erroneous had drawn attention according to the attention model. This performance was reached using attention model parameter settings $\alpha = 1.09 \cdot 10^{-4}$,
\( \gamma = 106.19 \) and \( \lambda = 0.82 \). The AUC \( \text{RQ1} \) of the model was 0.82 ("good" in the traditional academic point system) and was considerably higher than 0.5.

5.3. Match for Increasing View Interval Durations (RQ2)

Fig. 6 shows the AUC \( \text{RQ2} \) for different view interval durations. The fitted curve is clearly above the diagonal line (i.e., the line corresponding to a linearly increasing model match), specifically for smaller view interval durations: After 1 second (see coordinate a in Fig. 6), which is 3.5% of the average time the participants spent during view intervals (29 seconds), already 54.8% of the AUC \( \text{RQ2} \) at 29 seconds (far right value) was reached. After 3 seconds (10.3% of 29 seconds) 68.4% (b), after 9.8 seconds (33.8% of 29 seconds) 80% (c) and after 17.6 seconds (60.7% of 29 seconds) 90% of AUC \( \text{RQ2}(29) \) was reached (d).

Fig. 7 shows that the tuned parameters indeed depend on the used view interval duration. Consistent with this, Fig. 8 shows that tuning these parameters per view interval duration results in higher model performance (AUC of 7.55 for untuned versus 8.30 for tuned parameters). In general, \( \gamma \) peaks around seven seconds, while both \( \lambda \) and \( \alpha \) increase with view interval duration. This implies that for short view interval durations, attention threshold \( \alpha \) is best set relatively low and decay \( \lambda \) relatively high. This makes sense because in order to be able to pass a threshold in a short time, this threshold should be low, and in order to take advantage of the decay parameter, there should be some decay possible in a short time.
Figure 5: ROC curve for different optimal parameter settings of the model for varying hit rate settings, with the hit rate on the y-axis and the false alarm rate on the x-axis. Participants can be discerned in the dotted ROC curves. The blue curve depicts the average. The green cross indicates the point that is furthest away from the line of no-discrimination. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)
Figure 6: Match between modeled and marked geological model errors, expressed as $AUC_{RQ2}$ and percentage of the $AUC_{RQ2}$ (at 29 seconds) for increasing view interval durations. The solid blue line is the fitted curve, the dashed blue lines indicate the confidence interval. The red line corresponds to a linearly increasing model match. The points a, b, c and d are reference points. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)
Figure 7: View interval duration dependency of the three free attention model parameters $\gamma$ (distance from gaze), $\lambda$ (decay over time) and $\alpha$ (attention threshold).
Figure 8: Attention model match for increasing view interval duration with (higher solid blue line) and without (lower dashed green line) tuning per view interval duration. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)
6. Discussion and Conclusions

In this paper we took a first step to explore the potential of a new type of quality control of geological voxel models based on gaze. We explored what geologists regard as ‘erroneous’ by using their gaze and by implementing this information into a right-wrong estimate. We examined 1) to what extent attended areas as determined from experts’ gaze match with manually marked geological model errors, and 2) how this match improves over viewing time. To translate gaze data into attended areas, we used an attention model as described by Bosse et al. (2009, 2012). We found a clear match between manually marked errors and attended areas as determined using gaze. We also found that a large proportion of this match is reached within a small amount of total viewing time. For example, results show that after 3 seconds (i.e. 10% of the average viewing time) 68% of the match was reached. This is consistent with earlier studies that indicated that ‘intuition’ is used in nearly all human decision making and that it initiates quickly after a problem is detected (Kahneman, 2011).

While our results clearly indicate a match between manually marked and modeled errors, the match is not perfect. Obviously, a relatively simple attention model based on gaze alone will not capture attention with 100% accuracy. More importantly, a perfect match is not expected due to the following reasons: Firstly, attention does not equal error detection. Besides errors, attention is also likely drawn by areas that only look like errors due to salient visual features. Secondly, some discrepancy is expected between gaze and manually indicated errors due to (un)conscious ‘thresholds’ that geological experts will apply before marking certain areas as errors. Gaze
might therefore relate to other errors than those marked by hand. This means that most errors are found by using a combination of modeled and manually marked errors.

We may improve detection of geological model errors based on gaze by increasing our insight as to what exactly underlies the misses and false alarms. For example, for some misses we may find that, since geological model errors often occur at the same \((x,y)\) location for a sequence of slices, experts do not gaze at this location for yet another slice long enough to be marked as an error by the attention model. Or for some false alarms, we may find that attended areas constitute visually salient features, are geologically interesting or complex but not erroneous, look like geological model errors, or are in fact (or on second thought) geological model errors. The gaze-based attention model might then be improved by adapting attention thresholds of new slices based on previous slices, or on areas that are a priori expected to draw attention by using a saliency model like that of Itti and Koch (2000). An approach similar to the latter was proposed by Khosla et al. (2012). They selected salient parts of images that had to be searched for suspicious elements using saliency models that are based on low level imagery features such as color contrasts (Huber and Khosla, 2011). Blurring out these features could speed up the subsequent search by the human who now spends his or her time on searching for objects or events that require expert knowledge rather than being distracted by salient visual features. Another route of improving the model would be to add information from brain signals. It has been shown that by relating EEG traces to eye fixations, it is possible to distinguish above chance whether an observer is looking at a looked-for target (which in this
case would be a geological model error) or not (Brouwer et al., 2013, 2014; Kamienkowski et al., 2012).

The overarching challenge is to use our findings in future geological voxel model quality control workflows. Research should therefore focus on concrete implementation of facilitating and optimizing error detection in geological voxel models or in other types of models in need of (otherwise time-consuming) quality control, such as a deterministic layer model. Future research should also focus on the development and evaluation of interfaces that for instance allow experts to easily fix any detected model errors. Furthermore, it should be investigated whether presentation of different stimuli to experts (such as projected borehole data along the slices) enhances quality control. These improvements should then be compared to conventional quality control systems.

Our finding of rapid error detection at the beginning of viewing intervals suggests that an expert could perform first order quality control within 10% of the time they would normally use. Thus, model slices may be presented for relatively short view time interval durations to experts who are asked to only look for errors. Speeding up the process gains time that makes it possible to examine more data or to have different types of experts examine the same data. An analysis of the nature of the errors allows for independent comparison between different geologists with respect to their focus and expertise. This would facilitate a more effective use of experts in different projects or areas. Part of such a procedure could be to confront experts with the possible areas of interest as determined from their own gaze behavior. Similar to the approach of Krupinski et al. (1998), this allows for possible
training steps to improve geological expertise and eventually the geological voxel model itself.

Acknowledgments

We thank all participants for allowing us to collect the data. This research was funded by the TNO Enabling Technology Program “Models”.

References


