Remote Sensing Letters
Vol. X, No. X, Month 2010, xxx–xxx

Remote sensing meets psychology: a concept for operator performance assessment

SOETKIN GARDIN*†, SÉBASTIEN M.J. VAN LAERE‡, FRIEKE M.B. VAN COILLIE†, FREDERIK ANSEEL§, WOUTER DUYCK‡, ROBERT R. DE WULF† and LIEVEN P.C. VERBEKE¶

†Laboratory of Forest Management and Spatial Information Techniques, Ghent University, Coupure Links 653, 9000 Gent, Belgium
‡Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, 9000 Gent, Belgium
§Department of Personnel Management, Work and Organizational Psychology, Ghent University, Henri Dunantlaan 2, 9000 Gent, Belgium
¶Geo Solutions, Veldkant 33B, 2550 Kontich, Belgium

An often undervalued but inevitable component in remote sensing image analysis is human perception and interpretation. Human intervention is a requisite for visual image interpretation, where the interpreter actually performs the analysis. While image processing became more and more automated, human screening and interpretation remained indispensable at certain stages. One particular stage where the operator plays a crucial role is in the development of reference maps. This is often done by a visual interpretation of an image by one operator. Although the result is crucial for adequately assessing automated systems' performance, the work of the human operator is rarely questioned. No variability is considered and the possibility of errors is not mentioned. This is an implicit assumption that operator performance approaches perfection and that infrequent errors are randomly distributed across time, operators and image types. Given that the existence of operator variability has been proven in several other related domains, *e.g.* screening of medical images, this assumption may be questioned. This letter brings the issue to the attention of the remote sensing community and introduces a new concept quantifying operator variability. As the WAVARS project will gain from a high amount of data, we kindly invite interested researchers to access the website http://wavars.ugent.be and take part in the test.

1. Introduction

Remote sensing image analysis has become increasingly automated during the past decades. New algorithms are being developed that are more cost- and time-efficient than the traditional human-operated procedures (Zitova and Flusser, 2003). However, notwithstanding these technological developments, human intervention remains inevitable, even crucial, throughout the remote sensing image analysis process. Early in the process, when images are (co-)registrated, the operator is expected to accurately and precisely localize ground control points. Later, images are analyzed by means of algorithms which in turn require training data and parameter tuning, which are both fairly subjective operator tasks. Even highly automated photogrammetric processing chains or image fusion techniques are not fully operational without the intervention of human interpreters. Finally, when it comes to algorithm validation and accuracy assessment of the image products, a human operator intervention is again required.

* Corresponding author. Email: Soetkin.Gardin@UGent.be

Remote Sensing Letters
ISSN xxxx–xxxx print/ ISSN xxxx–xxxx online
http://www.tandf.co.uk/journals
DOI: 10/1080.xxxxxxxxxxxxx

A particular implicit assumption that is rarely questioned is that of the operator's performance accuracy. Instead, an evaluation of accuracy is typically conducted by comparing the result of some automated process with an operator segmentation/classification that is called "ground *truth*" (Carleer *et al*., 2005; Esch *et al*., 2008). As operator performance is considered as *truth*, assuming zero variance between judges, the human operator is implicitly assumed to be infallible. Foody (2002) questions this assumption by stating that an algorithm's accuracy assessment is actually only reflecting the degree of correspondence with these ground data, but not necessarily with reality.

2. Background

In recent years, several studies have started acknowledging human factors as a possible source for observed errors in remote sensing image analysis. Some researchers even explicitly focused on operator variability by comparing the work of multiple operators (e.g. Powell *et al*., 2004; Leckie *et al*., 2003; Zhou *et al*., 2010). As results strongly differed among operators, researchers were confronted with the subjectivity related to particular interpretation jobs. Secondly they observed a difference in variability level depending on the performed tasks. Apparently, some sort of unwritten consensus among (experienced) operators exists on how certain elements should be interpreted, while there is a lot of subjectivity involved in other, more difficult tasks (Zhou *et al*., 2010). Next to these studies that are specifically oriented towards the operator, variability is sometimes encountered rather unwillingly when researchers are examining why the accuracy assessment of an automated process didn't show the expected result, just to find that the latter was corrupted by errors present in the (visually interpreted) reference data. Congalton (1991) already warned for a careful use of photo-interpreted images as a reference in accuracy assessment with an error matrix: accepting the photo-interpretation to be correct without any confirmation might lead to a rather poor and unfair assessment of the digital classification.

Two strategies have been adopted in order to tackle the problem of errors originating in the reference map during visual interpretations.

A first approach is the use of multiple operators. For instance, Zhu *et al*. (2000) asked two operators, of whom one was familiar with the environment, to define boundaries on the same image. In cases of disagreement, a third operator decided what the final boundary would be. Approximately 30 percent of the sample sites had to be visited by all three analysts to resolve interpretation differences. This illustrates the frequency of interpretation ambiguities and shows that aggregating over operators may not always resolve problems.

A second group of studies addressed human performance variability by explicitly assessing interpreter coding confidence (e.g. Sarmento *et al*., 2009). A region with low accuracy should be associated with low confidence and errors in the reference map. Sarmento *et al*. (2009) noted that "the general confidence in the reference classification in this example tended to be low". This again illustrates the uncertainty inherent to visual interpretation (see also Gonçalves *et al*., 2009; Leckie *et al*., 2003; Scepan, 1999; Scepan *et al*. 1999).

Hence it can be stated that it has been sporadically acknowledged that human performance variability may affect remote sensing image analysis results. However, because no systematic research investigated this issue, it remains unclear to what extent research conclusions may be questioned, what the sources of variability are, and what strategies researchers/practitioners could adopt to mitigate this problem.

Another important operator related factor to be addressed in remote sensing studies is the existing variability in working strategy. This is in the first place caused by the level of experience leading operators to the application of more standard procedures. Hoffman *et al*. (2001) establishes the importance of experience as a starting point for problem solving strategies in remote sensing image analysis. But also between people with the same level of experience different strategies can be identified, for example in the use of additional data. Scepan *et al.* (1999) observed that even expert operators tend to make different choices when confronted with the possibility to use additional data during a land classification. Some operators collect as much data as possible, while others prefer to keep an open mind only using the presented images. Albrecht *et al*. (2010) observed that students (n=24) who were all assigned the same image interpretation task applied a different amount of generalization. It could be stated that this was caused by a lack of experience of the students forcing them to use their own intuition of what makes a good interpretation. On the other hand, in spite of close follow-up to obtain a homogeneous result two highly trained and instructed operators still reached a different level of detail (Madden *et al*., 2010), proving that a certain amount of human variability is inevitable.

This letter mainly focuses on errors and variability among operators, but studying operator variability can also lead to other insights. A couple of studies have focused on a completely different factor, namely the way operators perceive features (e.g. Hoffman *et al.*, 2001; Popple, 2003). According to Hoffman *et al.* (2001), better understanding of how humans perceive images could be the key to the development of better algorithms. In 1998, Hodgson already doubted the value of very commonly used window sizes in remote sensing image analysis as the adopted sizes were often too small for even an expert operator to make a decision, let alone that a machine could do it. The recent shift from pixel-based to object-based image analysis also proves the growing attention for methods that show similarities with the human interpretation process. Corcoran *et al.* (2010) even went a step further by using a cognitive experiment in their attempt to model the human visual process of primitive-object segmentation.

In line with Corcoran's research we suggest that remote sensing research might greatly benefit from established results from psychological research on signal detection theory. This approach has been particularly fruitful in scientific domains that impose similar demands on human operator as in remote sensing image analysis/interpretation. In most of these related fields, it is already known for several years that individual, psychological variables substantially affect visual scanning and vigilance performance (air traffic control: Bisseret, 1981; industrial inspection tasks: Drury, 1975; eyewitness testimony: Ellison and Buckhout, 1981; child abuse: Laming, 2004; medical diagnosis in cancer screening: Laming and Warren, 2000).

Such sustained attention studies typically imply tasks in which participants have to monitor a screen over a prolonged period of time for an unpredictable target stimulus or signal (for reviews, see Davies and Parasuraman, 1982; Parasuraman, 1986). Thus, in these psychological signal detection studies, human operators are passive observer of dials, video screens, and other sources of information (Kessel and Wickens, 1982), analogous to the routine interpretation of remote sensing imagery by humans. One of the key findings in these classical studies has been that operator performance drops drastically after 30 minutes. After several decades of research on this phenomenon, researchers have even concluded that this vigilance decrement is "about as dependable a result as one will ever see in human experimentation" (Wiener, 1987).

Given the high similarity between these situations and the long-lasting routine tasks that are typically required from operators involved in remote sensing image analysis, we propose that the insights from these classical attention studies should be utilized to examine, understand, and improve human performance in various image analysis procedures. Regrettably, in spite of the high Earth Observation research output during the last three decades, human factors that may affect operator accuracy have not yet been the focus of research efforts, nor do any objective instruments exist to evaluate the effect of operator performance on remote sensing image analysis. Therefore, we suggest establishing a systematic research agenda examining the role of human performance in remote sensing. This will be of considerable value to key remote sensing and GIS stakeholders, such as application developers and service providers in the field of land survey.

3. Research concept

As a first step, we hereby introduce the research project WAVARS (Web-based assessment of operator performance variability within remote sensing image interpretation tasks) that aims to provide more insight into operator performance and the nature of errors as a function of problem characteristics. These objectives will be examined through a series of large-scale, ecologically valid and mainly web-based experiments. A pilot experiment has already been run with university students and personnel. In order to ensure sufficiently large datasets, further data will be collected through the internet. Test persons will go through two series of tests. First, demographic characteristics, a psychological personality profile and measures of cognitive functioning (visual working memory: Luck and Vogel, 1997) will be assessed. Secondly, a long series of digitizing tasks with remote sensing images is presented. The results of these tasks will be compared with highly accurate reference data. These data are collected by (mobile) land survey and photogrammetric interpretation of high resolution images followed by a long series of accuracy checks. As the web application only offers low resolution aerial images, associated errors in the reference data are significantly smaller than operator inaccuracies in the digitizing tasks provided by the web application. The collected data will be used to assess (1) the expected time-related vigilance drop, (2) the effects of operator personality, within the Big-Five framework (McCrae and Costa, 1987; McCrae and John, 1992), (3) the effects of operator demographics (age: Giambra and Quilter, 1989; gender: Halpern, 1986; Kimura and Hampson, 1994) and (4) the effects of cognitive individual differences on sustained attention abilities during image analysis tasks. In addition, the data will be used to estimate the within-group variance of operator performance as a function of problem characteristics.

In a second step, we aim to develop strategies to improve human performance during remote sensing image analysis tasks. Human competences should not be considered static and operators should be given the opportunity to develop their skills. Given the ample evidence that feedback improves performance on vigilance tasks (Green and Bavelier, 2003), we will develop a feedback intervention that could be used to train individuals in image analysis tasks. This feedback intervention would consist of a number of typical remote sensing tasks that individuals can complete online ('simulations'). After a first trial, individuals will receive constructive feedback and complete an online feedback facilitation session designed to increase and maintain skill development. This feedback tool will be made freely available on the internet so that the remote sensing community can optimally benefit from the results of the WAVARS project.

Thirdly, the results of this study should allow researchers and stakeholders to identify and select human operators for remote sensing image analysis tasks. Starting from Rose *et al*. (2002), who found correlations between personality factors and performance on sustained attention measures, an optimal personality profile for tasks that require sustained attention will be identified. This profile should be useful to the development of an assessment instrument to identify and select individuals, who feature the appropriate knowledge, skills, and abilities to perform image analysis tasks, with a high level of accuracy for longer periods of time. The benefits of such an instrument are obvious: on the basis of this cost-effective instrument, research institutions, private and governmental organizations will be able to screen and identify individuals for remote sensing tasks, leading to more accurate outcomes. The developed instrument will be made available to the broad remote sensing community through the internet.

References

ALBRECHT, F., LANG, S. and HÖLBLING, D., 2010, Spatial accuracy assessment of object boundaries for object-based image analysis. In *Third international conference on all aspects of GEOGRAPHIC OBJECT-BASED IMAGE ANALYSIS*, 29 June - 2 July, Ghent, Belgium.

BISSERET, A., 1981, Application of signal detection theory to decision making in supervisory control. *Ergonomics*, **24**, pp. 81-94.

CARLEER, A.P., DEBEIR, O. and WOLFF, E., 2005, Assessment of very high spatial resolution satellite image segmentations. *Photogrammetric Engineering & Remote Sensing*, **71**, pp. 1285-1294.

CORCORAN, P., WINSTANLEY, A. and MOONEY, P., 2010, Segmentation performance evaluation for object-based remotely sensed image analysis. *International Journal of Remote Sensing*, **31**, pp. 617-645.

CONGALTON, R.G., 1991, A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, **37**, pp. 35-46.

DAVIES, D.R. and PARASURAMAN, R., 1982, *The Psychology of Vigilance* (London: Academic Press).

DRURY, C.G., 1975, Inspection of sheet materials: Model and data. *Human Factors*, **17**, pp. 257-265.

ELLISON, K. and BUCKHOUT, R., 1981, *Psychology and Criminal Justice* (New York: Harper & Row).

ESCH, T., THIEL, M., BOCK, M., ROTH, A. and DECH, S., 2008, Improvement of image segmentation accuracy based on multiscale optimization procedure. *IEEE Geoscience and Remote Sensing Letters*, **5**, pp. 463-467.

FOODY, G.M., 2002, Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, **80**, pp. 185-201.

GIAMBRA, L.M. and QUILTER, R.E., 1989, Sex differences in sustained attention across the adult life span. *Journal of Applied Psychology*, **74**, pp. 91-95.

GONÇALVES, L.M.S., FONTE, C.C., JÚLIO, E.N.B.S. and CAETANO, M., 2009, A method to incorporate uncertainty in the classification of remote sensing images. *International Journal of Remote Sensing*, **30**, pp. 5489-5503.

GREEN, C.S. and BAVELIER, D., 2003, Action video game modifies visual selective attention. *Nature*, **423**, pp. 534-537.

HALPERN, D.F., 1986, *Sex Differences in Cognitive Abilities* (Hillsdale, NJ: Erlbaum).

HODGSON, M.E., 1998, What size window for image classification? A cognitive perspective. *Photogrammetric Engineering & Remote Sensing*, **64**, pp. 797-807.

HOFFMAN, R.R., MARKMAN, A. and CARNAHAN W.H., 2001, Angles of regard: psychology meets technology in the perception and interpretation of nonliteral imagery. In: *Interpreting Remote Sensing Imagery: Human Factors*, Hoffman R.R. and Markman, A. (Eds.), Chapter 2, pp. 11-55 (New York: Lewis Publishers).

KESSEL, C.J. and WICKENS, C.D., 1982, The transfer of failure-detection skills between monitoring and controlling dynamic systems. *Human Factors*, **24**, pp. 49–60.

KIMURA, D. and HAMPSON, E., 1994, Cognitive pattern in men and women is influenced by fluctuations in sex hormones. *Current Directions in Psychological Science*, **3**, pp. 57-61.

LAMING, D. and WARREN, R., 2000, Improving the detection of cancer in the screening of mammograms. *Journal of Medical Screening*, **7**, pp. 24-30.

LAMING, D., 2004, *Human Judgment: The Eye of the Beholder* (London: Thomson Learning).

LECKIE, D.G., GOUGEON, F.A., WALSWORTH, N. and PARADINE, D., 2003, Stand delineation and composition estimation using semi-automated individual tree crown analysis. *Remote Sensing of Environment*, **85**, pp. 355-369.

LUCK, S. J. and VOGEL, E. K., 1997, The capacity of visual working memory for features and conjunctions. *Nature*, **390**, pp. 279-281.

MADDEN, M., JORDAN, T., MASOUR, J. and WANG, J., 2010, GEOBIA-based spatio-temporal changes and geovisualization within wildland urban interfaces. In *Third international conference on all aspects of GEOGRAPHIC OBJECT-BASED IMAGE ANALYSIS*, 29 June - 2 July, Ghent, Belgium.

MCCRAE, R.R. and COSTA, P.T.J., 1987, Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, **52***,* pp. 81-90.

MCCRAE, R.R. and JOHN, O.P., 1992, An introduction to the five-factor model and its applications. *Journal of Personality*, **60**, pp. 175-215.

POPPLE, A.V., 2003, Context effects on texture border localization bias. *Vision Research*, **43**, pp. 739-743.

POWELL, R.L., MATZKE, N., DE SOUZA JR. C., CLARK, M., NUMATA, I., HESS, L.L. and ROBERTS, D.A., 2004, Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. *Remote Sensing of Environment*, 9**0**, pp. 221-234.

PARASURAMAN, R., 1986, Vigilance, monitoring, and search. In: *Handbook of Perception and Human Performance. Vol. 2. Cognitive Processes and Performance*, Boff, K.R., Kaufman, L. and Thomas, J.P. (Eds.), Chapter 43, pp. 1-39 (Chichester: Wiley).

ROSE, C.L., MURPHY, L.B., BYARD, L. and NIKZAD, K., 2002, The role of the big five personality factors in vigilance performance and workload. *European Journal of Personality*, **16**, pp. 185-200.

SARMENTO, P., CARRÃO, H., CAETANO, M. and STEHMAN S.V., 2009, Incorporating reference classification uncertainty into the analysis of land cover accuracy. *International Journal of Remote Sensing*, **30**, pp. 5309-5321.

SCEPAN, J., 1999, Thematic validation of high-resolution global land-cover data sets. *Photogrammetric Engineering and Remote Sensing*, **65**, pp. 1051-1060.

SCEPAN, J., MENZ, G. and HANSEN, M.C., 1999, The DISCover validation image interpretation process. *Photogrammetric Engineering and Remote Sensing*, **65**, pp. 1075-1081.

WIENER, E.L., 1987, Application of vigilance research: Rare, medium, or well done? *Human Factors*, **29**, pp. 725–736.

ZHU, Z., YANG, L., STEHMAN, S.V. and CZAPLEWSKI, R.L., 2000, Accuracy assessment for the U.S. Geological Survey Regional Land-Cover Mapping Program: New York and New Jersey Region, *Photogrammetric Engineering and Remote Sensing*, **66**, pp. 1425-1438.

ZHOU, W.Q., SCHWARZ, K. and CADENASSO, M.L., 2010, Mapping urban landscape heterogeneity: agreement between visual interpretation and digital classification approaches, *Landscape Ecology*, **25**, pp. 53-67.

ZITOVA, B. and FLUSSER, J., 2003, Image registration methods: a survey. *Image and Vision Computing*, **21**, pp. 977-1000.