Interpretation, Evaluation and the Semantic Gap ... What if we Were on a Side-Track?

Bart Lamiroy

Interpretation, Evaluation and the Semantic Gap ... What if we Were on a Side-Track?

Bart Lamiroy

Université de Lorraine – LORIA (UMR 7503)
Campus Scientifique – BP 239
54506 Vandoeuvre-lès-Nancy CEDEX, FRANCE
Bart.Lamiroy@loria.fr

Abstract. A significant amount of research in Document Image Analysis, and Machine Perception in general, relies on the extraction and analysis of signal cues with the goal of interpreting them into higher level information. This paper gives an overview on how this interpretation process is usually considered, and how the research communities proceed in evaluating existing approaches and methods developed for realizing these processes. Evaluation being an essential part to measuring the quality of research and assessing the progress of the state-of-the art, our work aims at showing that classical evaluation methods are not necessarily well suited for interpretation problems, or, at least, that they introduce a strong bias, not necessarily visible at first sight, and that new ways of comparing methods and measuring performance are necessary. It also shows that the infamous *Semantic Gap* seems to be an inherent and unavoidable part of the general interpretation process, especially when considered within the framework of traditional evaluation. The use of Formal Concept Analysis is put forward to leverage these limitations into a new tool to the analysis and comparison of interpretation contexts.

1 Introduction

One of the very basic aspects of experimental research is the level of "*verifiability*" and traceability of claims an results published by their authors. In [1] we already wrote that the basics of reproducible research, set by Popper [2], notably

1. reporting of clearly set goals and defined interpretation framework,
2. full access to all experimental data,
3. reporting of the experimental apparatus, setup and protocol, in such a way that it becomes fully reproducible,
4. all parameters defining the data (if applicable) and those related to the experimental process.

are difficult to achieve in real life; especially when related to document analysis. We reported in [3] that:

> The goal of document image analysis is to achieve *performance* using automated tools that is *comparable* to what a *careful* human expert would achieve, or at least to do *better* than *existing algorithms* on the same *task*.
> Our use of terms like "performance," "comparable," and "better" indicate that there is an underlying notion of *quality* and therefore *measurement*. It suggests a controlled process that continually improves toward perfection. However, we also make mention of "careful" humans, "tasks," and "existing algorithms." While humans may believe themselves to be expert and careful when performing a task, there are situations where they unavoidably disagree [4,5,6,7], meaning that, at best, quality and improvement are subjective notions. It also strongly suggests that, depending on the task, measurements will differ, advocating again for multiple ways of measuring overall performance. [...] It is important to note, however, that shared datasets are only a part of what is needed for performance evaluation, and since research in document analysis is often task-driven, specific interpretations of a dataset may exist. [...] This most certainly does not affect the intrinsic quality of the underlying research, but it does tend to generate isolated clusters of extremely focused problem definitions and experimental requirements. [...] It is generally assumed that there is a single, unambiguous annotation in every case and that it is recorded correctly in the ground-truth. [...] Existing tools allow the user to indicate how he/she believes a document should be interpreted, but do little to help users understand differences in interpretations. Such differences might be called "errors" when there is a strong consensus about what constitutes the right answer. In many cases, however, there are legitimate differences of opinion [4,8] by various "readers" of the document, and these may differ from the *intention* of the author (which is usually hard or impossible to determine, although sometimes we can get access to it [9]).
> The bottom line is that although standard document collections exist, their annotations or "ground truth" may be specific, recorded in predetermined representations, incomplete or partially erroneous, while, on the other hand, there is a need to collect and manage annotations in ways that make it possible to construct more robust and general document analysis solutions.

These excerpts, although taken from publications considering the problem initially from the angle of document image analysis easily apply to broader machine perception research. They raise the following fundamental questions:

1. How can individual contributions to the state-of-the-art, solving machine perception problems, be objectively evaluated? Can they be compared to previous work? Can there be a set of measurable criteria establishing that it actually contributes to improving the state-of-the-art?
2. In how far are these contributions constrained to a specific context of use? What is a context of use? Can it be described, formalized or measured?

3. Is it actually possible to evaluate a contribution with respect to human perception performance? Does it make sense? What would be required to be able to do so?

While these questions seem to be naively simple and common sense, there does not seem to be any thoroughly established framework addressing them. They actually seem to be taken for granted and *"obvious"*. We shall prove in what follows that they are far from being so, and that considering them lightly actually leads to severely distorted perceptions of the quality of research in many ways.

In what follows we shall develop the following reasoning: first we analyze the way current machine perception research considers the definition and evaluation of interpretation problems and how it establishes the so called *Semantic Gap* [10] (Section 2); we conclude that uncertainty and ambiguity in ground truth is intrinsic to interpretation problems, and cannot be avoided in all but in the most trivial cases. Section 3 proposes a paradigm shift in the ways of measuring and modeling performance differences, by incorporating this intrinsic level of *difference of opinion* (rather than talking about *errors*) in interpretation by using Formal Concept Analysis.

2 Evaluation and the Semantic Gap

In this section we are taking a look into how the global research community is considering evaluation of interpretation problems. Many quite advanced and interesting benchmarking and evaluation initiatives exist, aimed at measuring the performance of machine perception methods (*cf.* all competitions at ICDAR 2013[1] as well as other initiatives like Pascal VOC [11], LSVRC[2], TREC [12], IMAGECLEF [13] ...) They all adopt the same meta-framework:

- Provide annotated training data (ground truth or golden standard)
- Develop perceptive interpretation algorithms that fit these training data
- Compare algorithms and rank them with respect to their performance on annotated test data, different from the training data.

The overall consensus is that, if the training data sufficiently covers all examples to be handled in a particular perception problem, this approach provides sufficient support for developing and evaluating appropriate solutions. The significant progress of artificial perception methods and applications in the last few decades seems to support this viewpoint. We partially reject these assumptions and claim that they are too rigidly biased toward the interpretation context fixed by the annotated data, that, consequently, the evaluation and ranking process is intrinsically flawed because of this, and that, eventually, this way of proceeding hinders the discovery and development of objectively quantifiable perception

[1] `http://www.icdar2013.org/program/competitions`
[2] `http://www.image-net.org/challenges/LSVRC/2012/`

approaches. It should be clear that this does not invalidate nor intends to disqualify current research methods. However, it tries to introduce complementary and theoretically supported metrics and methodology that lift the mentioned biases and shortcomings.

2.1 Analysis of the Performance Metrics

In general, the performance of methods is expressed with respect to their level of agreement to the *ground truth*. Often precision-recall based metrics are used to express this, but this does not need to be. Often, approaches also give a more detailed, multi-dimensional measure of performance [14].

The *ground truth* itself, and, subsequently, the coverage of the training data, is considered to be flawlessly representative of a well defined interpretation problem for which the state-of-the-art research is expected to provide an algorithm. It is interesting to take a look at what would happen if we accepted less than perfect ground truth. Let us assume that the ground truth were tainted with errors, and that $\varepsilon\%$ of it were annotated with wrong interpretations. This would mean that a method, achieving a 100% agreement with the ground truth, would actually be off the initial interpretation problem with $\varepsilon\%$, and that there is a non-null probability that any method within the $]100\%, 100 - \varepsilon\%]$ agreement range actually outperforms the one with the highest score. This is actually a well known problem, and has been studied before [15]. The only apparent solutions to this problem seem to be to either resort to very strict, but essentially human, verification and cross-verification of ground truth, to either use synthetically generated ground truth data, or to do away with classical metrics, altogether [16,17].

In the next section, we shall show that these apparent solutions are not sufficient, and that difference in interpretation is a core component of the overall interpretation problem.

2.2 Expressing the Interpretation Context

Independently of what was enunciated in the previous section, evaluating and comparing interpretation approaches fundamentally rely on *interpretation context*. The context is what defines the interpretation domain $\mathcal{I}$ as the set of possible interpretations, the application or input domain Δ of the data to be interpreted, and an oracle function $\mathcal{O}$ assigning the assumed-to-be correct interpretations to the input data:

$$\mathcal{O}(\Delta) = \mathcal{I}_{\mathcal{O}} = \{(\delta, \mathcal{O}(\delta))\}_{\delta \in \Delta}$$

It is generally assumed that *ground truth* is a sampling of $\mathcal{I}_{\mathcal{O}}$. Furthermore, it is often implicitly assumed (although rarely, if ever, actually formalized or factually established) that Δ is a manifold of some sorts and that some local continuity properties exist (generally related to tolerance to noise and small deformations of the input) such that it is likely that $\Delta|_i$ (*i.e.* the class of all data for which the oracle returns i) constitutes a sub-manifold of some kind. If sufficient sampling

points are chosen from $\Delta|_i$, the consensus is that appropriate techniques can provide a fairly accurate approximation of it. This is what underpins a large part of current Machine Perception research performance evaluation.

In this paper, we are not going to consider the question of whether the ground truth sampling (especially the training data) is representative and sufficient, and if it captures the complete scope of the interpretation context it is supposed to cover from an information theory, Shannon-Nyquist or linear algebra point of view, to name a few. While this is an essential and fundamental question, it obviously requires a much larger and elaborate study than what can be reported here. From here on, and for argument's sake, we are going to assume that the ground truth used in Machine Perception, is generally a sufficiently representative sampling for the intended interpretation context.

But what is this *interpretation context*? It is the set of rules, conditions and constraints that define whether a given interpretation $i \in \mathcal{I}$ applies to some given input data $\delta \in \Delta$. We previously associated this set to an oracle $\mathcal{O}$. We can reasonably assume there exists no known algorithm for $\mathcal{O}$, otherwise the corresponding Machine Perception problem would be solved (except, perhaps, for some performance issues). Although this sounds trivial, it actually leads to a very interesting paradox we are going to make explicit, here.

The Case of Human Annotated Ground Truth the most common approach to generating ground truth is to use human annotators. In this configuration, the annotators serve as instances of the oracle $\mathcal{O}$ and are provided with input data, for which they are to produce the corresponding interpretations, following clear instructions. These instructions correspond to the interpretation context and are defined as precisely as possible using both natural language and mathematically formalized criteria. The paradox arrises immediately: either the instructions are totally unambiguous, and identically interpreted by all human annotators; either the instructions are ambiguous at some point, and may create legitimate different interpretations, depending on the annotators' viewpoints. Yet, totally unambiguous, fully formalized and totally reproducible instruction sets bear a name: algorithms. Hence, if the interpretation context can be formalized, the Machine Perception problem is solved. Consequently, in the case of human annotated ground truth, it is impossible to avoid a certain level (may it be minimal) of ambiguity, and therefore legitimate differences of interpretation will persist.

This is actually supported by many findings. [6] reports an experiment of pixel-level human annotation for document binarization, for instance. [18], reporting on the Pascal Visual Object Classes (VOC) Challenge, spends a significant part of the paper on an account of the various conditions to acquiring, annotating and validating the data and refers to explicit annotation guidelines [19]. The annotation guidelines contain descriptions of the data like "*Bounding box should contain all visible pixels, except where the bounding box would have to be made excessively large to include a few additional pixels (<5%)*", "*Images which are poor quality (*e.g. *excessive motion blur) should be marked bad. However, poor*

illumination (e.g. *objects in silhouette) should not count as poor quality unless objects cannot be recognised.*" ... as for the categories, "Bus" includes minibus, and "Car" includes cars, vans, people carriers *etc.* but should not be labeled when only the vehicle interior is shown. Obviously, images like the one in Fig. 1 fall in an ambiguous category both whether they should be labeled as "Car" (as van) or "Bus" (as minibus) on the one hand, and whether they *only* depict the interior of the vehicle.

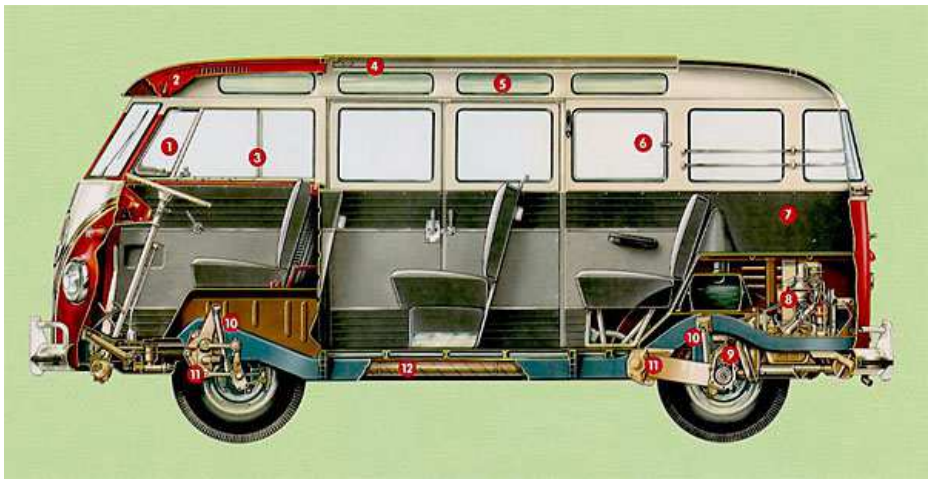

Fig. 1. Example of an ambiguous image category.
Source: `http://www.doobybrain.com/2008/01/30/car-cut-away-gallery/`

The Case of Synthetic Data synthetically generated ground truth is the dual configuration of human annotated ground truth, with respect to interpretation context. Indeed, in this case, there exists and algorithm $\mathcal{S}$ that is capable of generating data that is conforming to a given interpretation context. Formally speaking,

$$\begin{aligned} \mathcal{S} : \mathcal{I} \times \mathbf{P} &\rightarrow \Delta \\ i, p &\mapsto \delta \end{aligned} \tag{1}$$

where $\mathbf{P}$ is the parameter space of $\mathcal{S}$. Under those conditions, trying to determine an algorithm for $\mathcal{O}$ becomes an *Inverse Problem*, which is class of reputably hard, ill-posed problems, introducing a high level of ambiguity [20] in the general case.

It is interesting to consider the cases where $\mathcal{S}$ either is injective, surjective or bijective (other situations can, without loss of generality, be reduced to these three).

1. $\mathcal{S}$ is injective (and not bijective): this means that the generated ground truth does not cover the entire set of possible interpretation configurations, and therefore is not an appropriate, nor a representative tool for performance evaluation[3].
   Given that $\mathcal{S}$ can still be used for addressing a sub-part of the interpretation problem by restricting $\mathcal{O}$ to $\Delta' = \mathcal{S}(\mathcal{I}, \mathbf{P})$, the derived use comes down to considering the surjective or bijective case.
2. $\mathcal{S}$ surjective (and not bijective): this means that the interpretation problem is potentially ambiguous. If

$$\exists (i,p), (i',p') \in \mathcal{I} \times \mathbf{P} : i \neq i' \wedge \mathcal{S}(i,p) = \mathcal{S}(i',p')$$

   then there is a δ for which both interpretation i and i' hold[4]. However if, independently of any p, p'

$$\forall i \neq i' \in \mathcal{I} : \mathcal{S}(i,p) \neq \mathcal{S}(i',p')$$

   then the subjectivity is only due to an over-parametrization of the generative function, and has no impact on interpretation ambiguity. In that case the problem can be reduced, using an alternative $\mathcal{S}'(\mathcal{I}, \mathbf{P}')$, to the bijective case.
3. $\mathcal{S}$ is bijective: in that case $\mathcal{O} = \mathcal{S}^{-1}$.

Besides the fact that most of the synthetic ground truth generating methods have not been categorized into one of the above classes, and that, consequently, performance evaluation based on their use cannot be considered totally reliable (if not seriously flawed) they introduce a similar paradox as in the previous case: either the problem is well posed ($\mathcal{S}$ is bijective) but then it should be theoretically possible compute $\mathcal{O}$ as $\mathcal{S}^{-1}$ and the problem is solved by posing it; either the problem is ill-posed and any proposed solution will either be irrelevant ($\mathcal{S}$ is injective) or non-unique or ambiguous ($\mathcal{S}$ is surjective).

2.3 Standoff

The infamous Semantic Gap is here to stay, and seems to be a fundamentally intrinsic part of interpretation: either one is capable of very precisely state an interpretation problem, in which case the mere fact of stating it lifts any possible ambiguity and consists in solving it; either the problem is open to interpretation, and multiple contradictory solutions may fit the problem.

This is not really surprising, and is in line with post-modernist philosophic considerations on truth and interpretation [21,22]. While this does not mean that

[3] This is a somewhat strong statement, and in many cases it can be helpful to use these functions anyway, as an instance of common practice in experimental research: "If we cannot immediately solve the global problem, let's try and solve a more manageable sub-problem."

[4] We are making the implicit assumption that interpretations are mutually exclusive. Although this may seem restrictive, it is not. In cases where multiple interpretations are acceptable, one can simply replace $\mathcal{I}$ by $\{0,1\}^{|I|}$.

interpretation is impossible, it does conclude that multiple possible interpretations coexist and cannot be compared to one another. In the following section we shall be developing a set of computational tools to accommodate to this paradigm shift, very much in line with Eco's idea that only a limited number of all possible interpretations are worthwhile to consider [23,24].

3 Comparing and Modelling Differences in Interpretation Contexts

An extreme example of the previously mentioned standoff can be seen below. By considering for $\mathcal{I}$ set of allowable concepts $\{circle, triangle, square\}$, and for Δ the following input $\{\bigcirc, \triangle, \square\}$, Peirce's *unlimited semiosis* [22] would perfectly well admit the following interpretation

$$\bigcirc \models triangle$$
$$\triangle \models square$$
$$\square \models circle$$

Most of us would agree, however, that this interpretation is, at the least, unconventional, although one could imagine contexts where it actually makes sense (*e.g.* for obfuscation and cryptography[5]). The mere fact that conventional, self-imposing, interpretations exist (although they may not be unique) [23], hints that there may be common characteristics that can be extracted from them. In what follows, and in the light of the reasoning in the previous section, the term "algorithm" should be taken as equivalent to "interpretation context", and may therefore also refer to humans.

The main idea is to try and capture the possible structure underpinning the consensus and differentiation areas of a set of competing algorithms on the same data. By using Formal Concept Analysis [26,27] on the one hand, and possibly statistical clustering techniques on the other hand, we expect to be able to characterise their differences in interpretation.

The general idea is developed below.

We are assuming that the task at hand can be expressed as a discrete set of expected results $\mathcal{I} = \{i_1 \dots i_n\}$. If not, discretisation techniques like those developed in [28] can be used or adapted to fit the specific kind of descriptors used.

In that case, as in [17], we consider m algorithms $\{\mathcal{A}_k\}_{1\dots m}$ and a data set $\Delta = \{\delta_1 \dots \delta_d\}$. This allows us to construct a family of m $d \times n$ matrices M_k such that

$$M_k(s,t) = \begin{cases} 1 \textit{ iff } \mathcal{A}_k(\delta_s) = i_t \\ 0 \text{ otherwise} \end{cases}$$

Let the final matrix $\overline{M}$ be the concatenation of the matrices M_k such that $\overline{M} = [M_1 \dots M_m]$ as represented in Tab. 1.

[5] This fuzzy distinction between syntax, semiosis and semantics is actually what troubled interpretation of hieroglyphs [25]

	M_1				M_2				...	M_m			
	$\mathcal{A}_1$				$\mathcal{A}_2$				...	$\mathcal{A}_m$			
	i_1	i_2	...	i_n	i_1	i_2	...	i_n		i_1	i_2	...	i_n
δ_1	0	1		0	1	0		0		0	1		0
δ_2	1	0		0	0	0		1		0	1		0
$\vdots$													
δ_d	0	1		0	1	0		0		0	0		0

Table 1. Representation matrix $\overline{M}$ for FCA-based ground-truth interpretation and performance evaluation

By conducting a Formal Concept Analysis on these data, the appropriate clusters of coherent interpretations can be uncovered and compared with the "natural" concepts underpinning them. This will eventually result in a better understanding of how Machine Perception methods compare to one another in a more semantic sense, since the result of the FCA is a lattice structure, capturing the partial order (or hierarchy) of data/algorithm clusters sharing the same interpretation/agreeing. The interesting side-effect associated to this approach, which we also already discovered in [17], is that it contains a duality between data and methods, in the sense that it cannot only be seen as a tool for comparing and studying different algorithms, but that it can also be considered as a way to assess the appropriateness of data with respect to the methods.

The outputs we can use from the FCA are:

- For a given set of algorithms $\{\mathcal{A}_i\}$, which are the δ_k on which they share the same interpretation?
- Dually, for a given set of $\{\delta_k\}$, which interpretations i_j are observed and by which algorithms?
- Given a set of algorithms $\{\mathcal{A}_i\}$, what sets of disagreement of $\bar{\delta}_k$ exist, and how are they structured?
- What items $\bar{\delta}_k$ offer the largest level of disagreement (highest scattering between different observed interpretations)?

This last point is particularly interesting, since it offers a mathematically formalised metric for ambiguous data (in the case of [6], for instance, it allows to precisely identify the pixels for which the notion of binarisation does not seem to make much sense, or, at least, for which very legitimate differences of opinion exist). By extension, it offers a clearly defined bootstrap to extend the formalisation of existing interpretation contexts by precisely highlighting those input data on which it seems ambiguous. Combining this with current statistical classification methods may actually provide very interesting learning approaches, since they would focus on pertinent data.

Example

The ideas expressed in the previous section will require a profound paradigm shift with respect to comparing results from various sources. As an example,

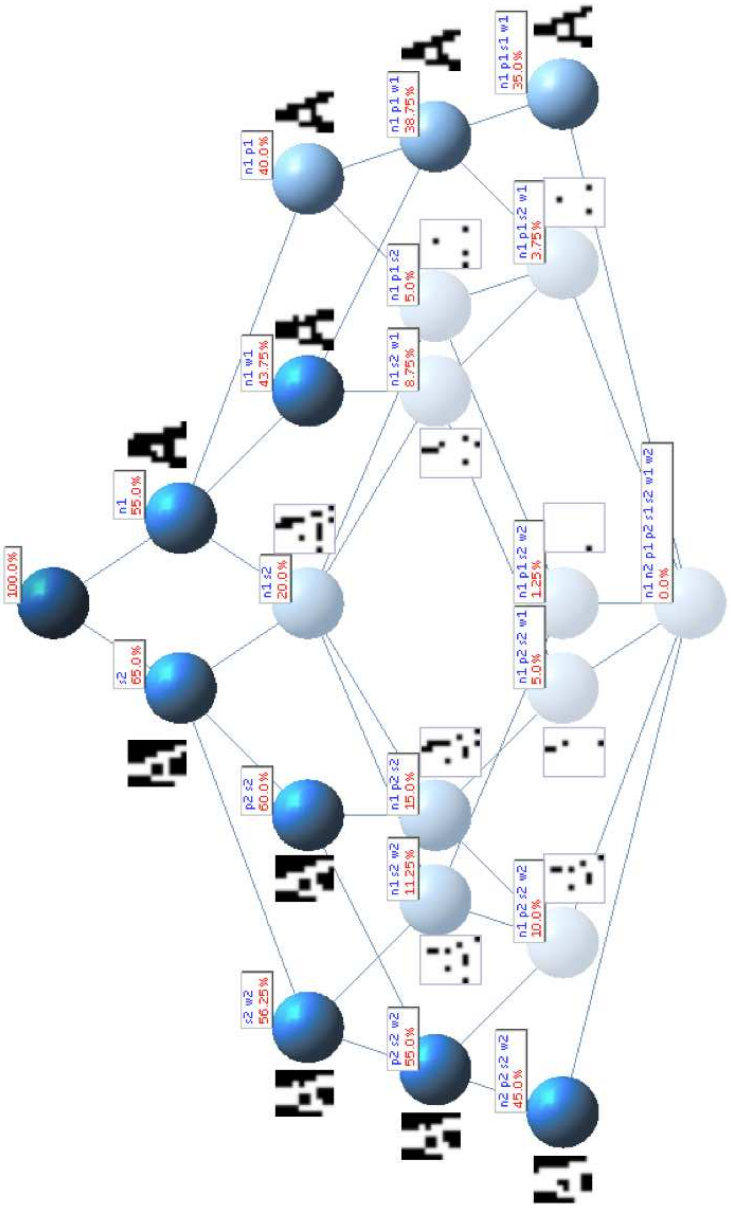

Fig. 2. Preliminary results on the use of FCA for binarisation algorithm comparison. Otsu (p) [29], Niblack (n) [30], Sauvola (s) [31] and Wolf (w) [32]. Suffix 1 to the algorithm names signifies foreground categorisation, suffix 2 means background categorisation.

Fig. 2 shows some preliminary results we obtained[6] by taking 4 off-the-shelf binarisation algorithms (Otsu [29], Niblack [30], Sauvola [31] and Wolf [32]) and by considering each pixel of a grey level image of an 'A' and feed their classification results to *Lattice Miner*[7] [33]. It needs to be noted that, unlike what is usually done in FCA, both attributes (here foreground pixels) and their complements (background pixels) have been used for concept construction. This creates a left-right symmetry in the concept lattice: the left side concerns concepts based on background information, the right side concerns foreground information. Our concept lattice that expresses the following knowledge[8].

1. Each concept node lists three elements: an image of the pixels belonging to the concept, the list of binarisation algorithms that *agree* on the categorisation of these pixels, the percentage of the whole image these pixels present.
2. The lattice hierarchy (top-down) goes in increasing order of combination of algorithms, and in decreasing order of categorised pixels. This means that a child node shows the pixels categorised by the parent node's algorithms that are also categorised by a another algorithm, which is added to the list of algorithms. In other terms, if a parent node shows the agreement of a set of algorithms, a child node shows the agreement of an extended set of these algorithms.

Some of the immediate conclusions that can be drawn from the output is that Niblack is consistently more optimistic than the others in classifying foreground pixels (to the point that any foreground pixel classified by the other approaches is also classified as such by Niblack), and conversely, that Sauvola is consistently more pessimistic than the others (all foreground pixels classified by Sauvola are also classified as such by the others). This is confirmed by the lattice structure on the left hand side, expressing the dual configuration on background pixels. Other interesting findings are in the center of the lattice, where it can be seen that consensus on foreground pixels between Niblack and Otsu on the one hand and the consensus on background pixels between Sauvola and Wolf on the other hand, are consistent with each-other, to the exception of a single pixel.

4 Conclusion

We have shown that ground truth is the instance of a very specific and unique interpretation context that is either immediately transposable into an algorithm (and therefore addressing an already solved problem) or otherwise fundamentally ambiguous. Not taking into account this intrinsic ambiguity has an impact on traditional performance evaluation consisting in measuring agreement/disagreement with ground truth, since there is no way of establishing

[6] Results by Z. Jiang, M.Eng. student at Mines Nancy, France.

[7] `http://sourceforge.net/projects/lattice-miner/`

[8] The reader should take into account that Fig. 2 has been rotated by 90^o. Top-bottom in the descriptive text translates to left-right in Fig. 2.

whether disagreements are due to incorrect implementations or caused by an alternative, legitimate interpretation due to this ambiguity.

In order to formally establish and measure differences in interpretation context, we propose to rely on Formal Concept Analysis, which is capable of computing a lattice structure that links, in a dual way, interpreted data and the interpreting algorithms (or humans) such that clusters of agreement and disagreement can be clearly established. Analysis of these clusters can then further determine to what extent algorithms are comparable on the one hand, and what categories of data are to be considered as ambiguous for a given set of contexts, on the other. Some preliminary results were shown on binarisation algorithms, but those can be extended to any other kind of interpretation problem.

References

1. Lamiroy, B., Lopresti, D.: An Open Architecture for End-to-End Document Analysis Benchmarking. In: 11th International Conference on Document Analysis and Recognition - ICDAR 2011, Beijing, Chine, IEEE Computer Society (2011) 42–47
2. Popper, K.R.: The Logic of Scientific Discovery. Reprint edn. Routledge (October 1992) Original edition, 1934 "Logik der Forschung".
3. Lamiroy, B., Lopresti, D.: A Platform for Storing, Visualizing, and Interpreting Collections of Noisy Documents. In: Fourth Workshop on Analytics for Noisy Unstructured Text Data - AND'10. ACM International Conference Proceeding Series, Toronto, Canada, ACM (2010)
4. Hu, J., Kashi, R., Lopresti, D., Nagy, G., Wilfong, G.: Why table ground-truthing is hard. In: Proceedings of the Sixth International Conference on Document Analysis and Recognition, Seattle, WA (September 2001) 129–133
5. Lopresti, D., Nagy, G., Smith, E.B.: Document analysis issues in reading optical scan ballots. In: Proceedings of the 8th IAPR International Workshop on Document Analysis Systems, Boston, MA, USA, ACM (2010) 105–112
6. Smith, E.H.B.: An analysis of binarization ground truthing. In: Proceedings of the 8th IAPR International Workshop on Document Analysis Systems, Boston, MA, USA, ACM (2010) 27–34
7. Clavelli, A., Karatzas, D., Lladós, J.: A framework for the assessment of text extraction algorithms on complex colour images. In: Proceedings of the 8th IAPR International Workshop on Document Analysis Systems, Boston, MA, USA, ACM (2010) 19–26
8. Lopresti, D., Nagy, G.: Issues in ground-truthing graphic documents. In: Proceedings of the Fourth IAPR International Workshop on Graphics Recognition, Kingston, Ontario, Canada (September 2001) 59–72
9. Eco, U.: The limits of interpretation. Indiana University Press, Bloomington (1990)
10. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12) (2000) 1349–1380
11. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html
12. Voorhees, E., Harman, D., et al.: TREC: Experiment and evaluation in information retrieval. Volume 63. MIT press Cambridgeˆ eMA MA (2005)

13. Mller, H., Clough, P., Deselaers, T., Caputo, B.: ImageCLEF: Experimental Evaluation in Visual Information Retrieval. 1st edn. Springer Publishing Company, Incorporated (2010)
14. Dosch, P., Valveny, E., Fornes, A., Escalera, S.: Report on the Third Contest on Symbol Recognition. In Wenyin Liu, J.L., Ogier, J.M., eds.: Graphics Recognition. Recent Advances and New Opportunities. Volume 5046 of Lecture Notes in Computer Science. Springer (2008) 321–328 French Techno-Vision program (Ministry of Research) Spanish project TIN2006-15694-C02-02 Spanish research program Consolider Ingenio 2010:MIPRCV (CSD2007-00018).
15. Carlotto, M.J.: Effect of errors in ground truth on classification accuracy. International Journal of Remote Sensing **30**(18) (2009) 4831–4849
16. Lopresti, D.P., Nagy, G.: Adapting the turing test for declaring document analysis problems solved. In Blumenstein, M., Pal, U., Uchida, S., eds.: Document Analysis Systems, IEEE (2012) 1–5
17. Lamiroy, B., Sun, T.: Computing Precision and Recall with Missing or Uncertain Ground Truth. In Kwon, Y.B., Ogier, J.M., eds.: Graphics Recognition. New Trends and Challenges. 9th International Workshop, GREC 2011, Seoul, Korea, September 15-16, 2011, Revised Selected Papers. Lecture Notes in Computer Science. Springer (2013) 149–162
18. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2) (June 2010) 303–338
19. Winn, J., Everingham, M.: The pascal visual object classes challenge 2007 (voc2007) annotation guidelines. `http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/guidelines.html` (2007)
20. Tarantola, A.: Inverse problem theory and methods for model parameter estimation. Society for Industrial Mathematics (2005)
21. Heidegger, M.: Being and Time. Library of philosophy and theology. Blackwell (1967)
22. Peirce, C.S.: Syllabus: Nomenclature and Division of Triadic Relations, as far as they are determined. MS [R] 540. (1903)
23. Eco, U., Collini, S., Culler, J., Rorty, R., Brooke-Rose, C.: Interpretation and Overinterpretation. Tanner Lectures in Human Values. Cambridge University Press (1992)
24. Eco, U.: Dall'albero al labirinto: studi storici sul segno e l'interpretazione. Bompiani (2007)
25. Champollion, J.: Précis du système hiéroglyphique des anciens égyptiens, ou recherches sur les élémens premiers de cette écriture sacrée, sur leurs diverses combinaisons, et sur les rapports de ce système avec les autres méthodes graphiques égyptiennes. Imprimerie royale (1828)
26. Ganter, B., Wille, R.: Formal concept analysis - mathematical foundations. Springer (1999)
27. Ganter, B., Stumme, G., Wille, R., eds.: Formal Concept Analysis, Foundations and Applications. In Ganter, B., Stumme, G., Wille, R., eds.: Formal Concept Analysis. Volume 3626 of Lecture Notes in Computer Science., Springer (2005)
28. Coustaty, M., Bertet, K., Visani, M., Ogier, J.M.: A new adaptive structural signature for symbol recognition by using a galois lattice as a classifier. IEEE Transactions on Systems, Man, and Cybernetics, Part B **41**(4) (2011) 1136–1148
29. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics **9**(1) (January 1979) 62–66

30. Niblack, W.: An Introduction to Digital Image Processing. Strandberg Publishing Company, Birkeroed, Denmark, Denmark (1985)
31. Sauvola, J., Pietikäinen, M.: Adaptive document image binarization. Pattern Recognition **33**(2) (2000) 225–236
32. Wolf, C., Jolion, J.M., Chassaing, F.: Text Localization, Enhancement and Binarization in Multimedia Documents. In: Proceedings of the International Conference on Pattern Recognition. Volume 2. (2002) 1037–1040
33. Lahcen, B., Kwudia, L.K.: Lattice miner: A tool for concept lattice construction and exploration. In: Supplementary Proceeding of International Conference on Formal concept analysis (ICFCA'10). (2010)