A Framework for the Assessment of Text Extraction Algorithms on Complex Colour Images

A. Clavelli, D. Karatzas and J. Lladós
Computer Vision Centre
Universitat Autònoma De Barcelona
Edifici O, Bellaterra, Spain
{aclavelli, dimos, josep}@cvc.uab.es

ABSTRACT
The availability of open, ground-truthed datasets and clear performance metrics is a crucial factor in the development of an application domain. The domain of colour text image analysis (real scenes, Web and spam images, scanned colour documents) has traditionally suffered from a lack of a comprehensive performance evaluation framework. Such a framework is extremely difficult to specify, and corresponding pixel-level accurate information tedious to define. In this paper we discuss the challenges and technical issues associated with developing such a framework. Then, we describe a complete framework for the evaluation of text extraction methods at multiple levels, provide a detailed ground-truth specification and present a case study on how this framework can be used in a real-life situation.

Categories and Subject Descriptors
H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *performance evaluation (efficiency and effectiveness).*

General Terms
Algorithms, Measurement, Performance.

Keywords
Performance Evaluation, Text Extraction, Colour.

1. INTRODUCTION
Over the last couple of decades, as the use of colour in publications has become affordable, a massive amount of colour paper documents have been produced (magazines, leaflets, posters, advertisements etc). At the same time, the traditional definition of document has been extended to incorporate any type of text container, including born-digital images (Web images, images in spam emails, CAPTCHAs), camera based images and video frames to mention just a few (see Figure 1). These new paradigms mark a departure from the traditional document analysis techniques as the omnipresence of colour and complex design schemes demand different approaches.

DAS'10, June 9–11, 2010, Boston, MA, USA.

(a) **(b)**

Figure 1. (a) Scanned poster, (b) Web and spam images.

Quite a few algorithms have been described in the literature for the location and extraction of text from such complex colour images [1][2][3][4][5]. Nevertheless, due to the considerable problems arising, the performance evaluation of such methods remains a grey area. In many cases, the authors report visual inspection results due to the lack of ground-truthed datasets and corresponding performance evaluation methods [4][5]. In other cases they chose an indirect performance evaluation by reporting character recognition results on the extracted image parts [3]. Any shortcomings of the OCR system (and there are usually a lot given the nature of such images) are reflected in the final results rendering such evaluation schemes unable to assess the text extraction part of the method on its own right.

There are different paths taken in the literature towards text extraction from complex colour images. Top-down approaches, usually based on texture analysis, aim to locate text zones in the image as opposed to the pixel-level separation of text from the background [6]. Performance evaluation in such cases is typically done by measuring the bounding box overlapping between the result and the ground truth [7][8] and is rooted to earlier evaluation schemes originally conceived for layout analysis algorithms [8][9][10]. This kind of evaluation is open to many problems, as pointed out in [1]. Although locating zones of text makes some sense in the case of under-sampled text where segmentation at the pixel level is difficult, further processing is needed before recognition is achieved. Attempts of word-spotting in the extracted zones or even the use of handwriting recognition techniques (e.g. HMMs) are possible [11] but generally inefficient.

Another set of methods work in a bottom-up fashion and aim to produce a pixel-level segmentation of the image so that different

characters are represented by separate connected components [5]. Nevertheless, by text extraction it is meant not only that segments are produced, but that they are actually classified as text or background. This is a problem on its own right, distinct from segmentation. A performance evaluation framework should be able to assess both the efficiency of the segmentation (separation of text from non-text areas at the pixel level) as well as the post-processing steps towards text extraction (connected component labelling, character restoration, grouping into words, etc).

In [1] the authors propose a text extraction method that combines pixel level segmentation with subsequent component filtering and component grouping into text zones. Nevertheless, for performance evaluation the authors have to revert to bounding box overlapping measures that are really assessing only the text zone locating performance. In [2] the authors propose a method for segmenting and grouping character-like components into text lines based on the creation of multiple hypotheses and text line ranking. The final results are assessed by comparing the set of pixels of the extracted components to manually created ground-truth binary images. This permits the calculation of global segmentation metrics, but no evaluation of the text line extraction part, which is the real objective here, is possible.

Existing performance evaluation frameworks do not generally work with pixel-accurate ground truth and are not suitable for this application. Attempts to define a pixel-accurate ground truth and performance evaluation frameworks, although advantageous present a lot of problems (see [12] for a nice discussion). Probably the most adequate framework for a pixel-level evaluation is the one proposed by Ntirogiannis el al [13] for the purpose of evaluating thresholding results. Some interesting ideas are indeed put forward, which are discussed in more detail in section 2.

A unified framework for the performance evaluation of text extraction methods should be able to evaluate such methods at multiple levels: pixel-level segmentation, character restoration, text localisation, word and text-line extraction. Such a framework, to the knowledge of the authors, does not yet exist.

This paper describes first our findings in the process of creating a comprehensive ground truth specification and corresponding performance evaluation framework for text extraction methods. The resulting framework is described next in sections 3 and 4. Following that, in section 5, a case study is presented illustrating the behaviour of the framework in real life situations. Finally, a discussion is given in section 6 and conclusions in section 7.

2. BACKGROUND

One of the key difficulties in developing an objective evaluation framework for text extraction algorithms in a colour image context lies with the inherent difficulties of specifying adequate ground truth information. Creating ground truth information typically involves tedious manual work. This is especially true here, as ground truth should be defined in different levels, including the pixel level.

Apart from the considerable manual effort required, the most challenging aspect when it comes to colour text images is to reconcile differences in interpretation by different observers. Locating and reading text is a high-level cognitive task, and prior knowledge and experiences play an important role in the visual perception of text.

(a) **(b)**

Figure 2. (a) Broken-by-design, (b) merged-by-design text.

Take for example the relatively simple (in terms of colour content) images of Figure 2. A human observer can easily see 3 characters comprising the word “IBM”. To this human observer, the fact that each character comprises many parts does not hinder recognition. Should we penalise a text extraction method if it is not able to group together all the constituent parts of a single character? Or, to take it to the extreme, the white area between the blue stripes should be part of the background or part of the character in the ground truth?

The opposite can be seen in the case of Figure 2(b). Should we penalise a text extraction algorithm for suggesting that the letters “oca” form a single component? They were definitely produced as such. If they were to be annotated as separate characters in the ground truth, where should the break be? When asking human observers to define ground truth information in such images, the agreement between observers stops at a very high level interpretation (there are four characters in the word “Coca”), but if lower level information is required (which pixels belong to each character) chances are that human observers will disagree.

An easy solution to this kind of problem is to label pixels as text or non-text ignoring their particular membership to a higher level concept (character, word, etc). This is for example the working principle in Ntirogiannis et al [13] who suggest a comprehensive scheme intended for the evaluation of thresholding methods. The scheme of [13] is probably the most adequate evaluation scheme suggested in literature that could potentially be used with complex colour images. Their scheme is based on a dual ground truth representation where an inner skeleton and an automatically produced corresponding connected component are used to calculate global Precision (number of false alarms and deformation pixels) and Recall (number of broken text and missing text pixels). The automatically computed connected components are calculated through the application of Canny edge detection on the greyscale image and a repetitive dilation of the (automatically computed and manually corrected) skeletons. Obviously in the case of complex colour images most of the automation introduced in [13] such as the calculation of the initial skeletons and the application of Canny edge detection cannot be applied and more manual work would be required.

(a) **(b)** **(c)**

Figure 3. (a) Ground Truth. (b) Dilated result and (c) deformed result presented over the GT.

The main issue though with the above scheme is that since global measures are reported (number of pixels), it is difficult to produce a qualitative evaluation of a method. See for example the case of Figure 3, the two results (b) and (c) would obtain the same scores in terms of the global metrics defined in [13] but there is a clear

qualitative difference between the two, as (b) preserves the morphological properties of the ground truth, and could potentially be recognised as the correct letter, and (c) does not. It is therefore more important for the performance evaluation framework to be able to measure the *degree to which morphological properties of the text are preserved* by the text extraction method than simply the number of pixels misclassified.

At the same time it is important not to loose sight to the fact that recognition is the final objective, hence we should be able to perform evaluation at a level higher than that of pixels. The question we are called to answer is "if OCR was perfect what percentage of text would be recognised given some text extraction results". In [12], where a similar pixel-level truthing problem is discussed, the authors state "considering any form of text, the next level up from pixel accurate ground truth would be at the character level, then the text line level and finally the paragraph level". Having examined the situations of *broken-by-design* and *merged-by-design* text exemplified in Figure 2, we would argue that the above statement is not necessarily true. The next level up from pixel accurate ground truth would be *text parts*, which would technically correspond to connected components.

The next level up from text parts is not necessarily characters, but has to do with the way text was produced. We will use here the term "*atom*" to refer to the minimum unit of production that can be recognised on its own. Atoms might comprise many text parts but still correspond to one character (e.g. the letter "I" of IBM) or comprise a single text part but correspond to many characters (e.g. the "oca" part of the Coca Cola logo). The reason for introducing the atom concept is that it is the smallest construct to which different observers would agree on (there is no common agreement on where exactly to split the "oca" text, but all would agree that it was produced as a single unit), while at the same time it is the smallest construct that contains enough information to be recognised (the parts of the letter "I" on their own mean nothing, and if one part is missing the character is severed). The next level up would be words, text lines and paragraphs although in the case where an atom spans two words the next level up from atoms would be directly that of a text line etc.

Armed with the idea of a dual representation introduced in [13], the necessity to measure the morphological properties of extracted text and the concept of atoms to deal with the *Broken-by-design* and *Merged-by-design* cases, we are almost ready to define a performance evaluation scheme. There is nevertheless one more important issue to discuss that is inherent to the fact that digital images (the focus of text extraction algorithms) are just samples of the ideal text content (the conceptual model humans utilise to interpret them, and inadvertently to ground truth them).

(a) **(b)**

Figure 4. (a) Camera captured text, (b) Anti-aliasing in a born-digital image.

In scanned documents the resolution is generally adequate to be able to define purely text and purely non-text pixels, but in digital-born and camera-based text images, there are a lot of cases where the exact border of characters is ill-defined and pixels cannot be clearly assigned to one of the two classes. See for example Figure 4. Enough pixels exist that clearly belong to the text area, so a skeleton can be defined for the text, but the extent of the characters is not easily identified. The expectation that ground-truthers will be able to define the text extents only approximately has to be taken into account in the performance evaluation framework.

The extreme case of the above example is text which is clearly there for the human observer but severely under-sampled, so that no skeleton can be defined. An example of this situation is given in Figure 5. In such cases, pixel-level ground-truthing (or pixel-accurate text extraction results) makes no sense and the framework should be able to indicate this, and provide the flexibility to define ground-truth information at higher levels only.

(a) **(b)**

Figure 5. (a) Web banner with under-sampled text, (b) detail of under-sampled text.

Taking into account all the above observations, we defined a ground truth specification (explained next) and a performance evaluation framework (explained in section 4) that is able to deal with the great majority of cases that complex colour images present.

3. GROUND TRUTH SPECIFICATION

The ground truth specification defined here takes into account the key aspects described above. A multi-level representation is therefore produced where the smallest entity described is a text part, followed by atoms (which might comprise just a single text part), words (a set of atoms) and text lines (a set of words or atoms). The inclusion of a paragraph level is trivial, but for the types of images considered here it is only rarely that a paragraph of text exists and is important. For simplicity we do not discuss the paragraph level here, although its definition would be the same as for words and text lines. A visualization of the different levels of the ground truth information for the poster of Figure 1a is shown in Figure 6.

For the basic element, the text part, a dual representation at the pixel-level is introduced following the ideas proposed in [13]. The representation consists of a skeleton and a corresponding area (see Figure 6a). The skeleton defines the minimum set of pixels that have to be identified for the text part to be considered morphologically intact, while the area defines the extents of the text part and is used, with certain elasticity, to check for deformations that could alter the morphology of the text part.

Atoms are defined as groupings of text parts as shown in Figure 6b. Atoms are further labelled based on their production complexity as *Normal*, *Broken-by-design* (atoms comprise many text parts), *Merged-by-design* (atoms comprise a single text part that corresponds to many characters) or both *Broken & Merged-by-design* (atoms comprise many text parts, some of which represent more than one character). The higher level entities (words and text-lines) are defined as groupings of lower-level

entities as shown in Figure 6c-d. In addition to the IDs of their comprising lower-level entities, bounding polygon information is also calculated for atoms, words and text-lines.

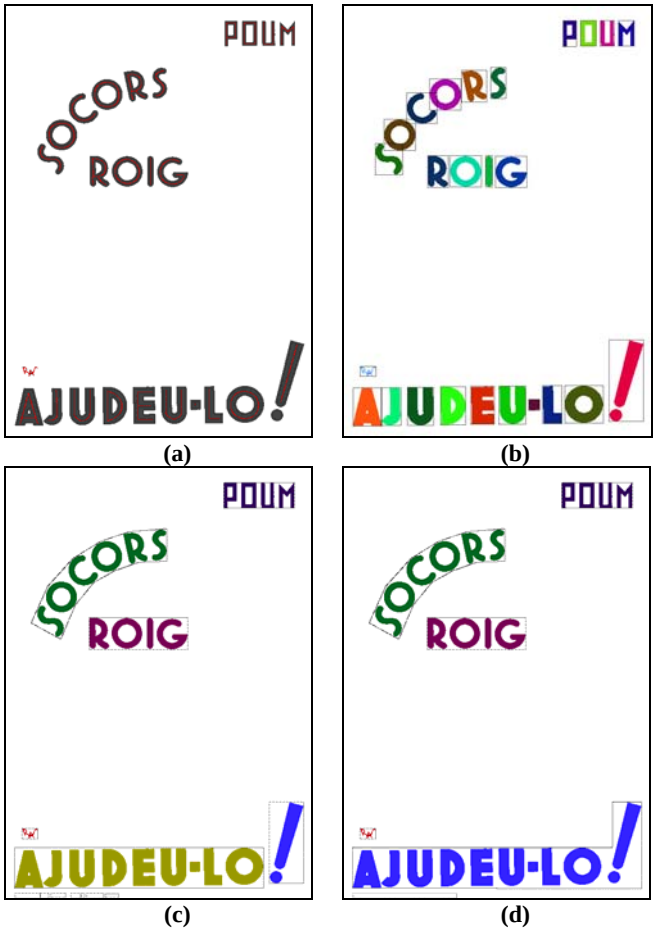

Figure 6. Visualizations of the ground truth information at different levels, (a) text part, (b) atom (c) word, (d) text line.

In the case of under-sampled text, where pixel-accurate labelling is not possible, the ground truth includes only bounding polygon information starting from the lower level possible. For example, the ground truth for the under-sampled text at the bottom-left of the poster image starts with the bounding box of words (see Figure 6c) and continues to define a text-line comprising these words (see Figure 6d). The above structure is defined in an XML schema[1]. A semi-automatic piece of software has been developed to permit the efficient definition of ground truth information at different levels, and the production of the appropriate XML file.

4. PERFORMANCE EVALUATION FRAMEWORK

Performance evaluation can take place in multiple levels according to the flavour of the text extraction method under evaluation.

The minimum expected output of a text extraction method is a set of atoms (which in the most basic case is just a list of connected components). Starting with a set of atoms two aspects of the text extraction method can be evaluated: its segmentation capability, namely the quality of text / non-text area separation at the pixel level and its atom extraction capability, namely the quality of the text part labeling and grouping process.

In the case of text zoning algorithms, or when pixel-level information is not possible to obtain, bounding box information is expected instead. In such cases the text localization capability of the method can be assessed, based on existing bounding box overlapping based schemes.

4.1 Assessment of segmentation performance

At this level, the objective is to assess the capability of the method to perform segmentation: pixel-level separation of text from non-text areas. The quality of the segmentation has to be assessed according to the suitability of the segmentation results for subsequent recognition.

Different algorithms can produce segmentations in which text appears a bit eroded or thickened while its overall shape is still preserved, or they may delete or add pixels to the text parts completely changing their morphology. Therefore the question we need to answer at this point is not *how many* pixels have been correctly labelled, but *which* pixels (see again Figure 3). To put it differently, we need to assess if an adequate number of pixels of each text-part have been correctly segmented so that the corresponding atoms are morphologically intact and can be recognised. Ultimately, we are interested in how many of the ground truth defined atoms were W*ell segmented* or not.

When assessing segmentation performance, we treat the output of the method under evaluation at its basic level which is a just a set of connected components. We ignore any grouping into atoms that the text extraction method has performed, as this is assessed in the next level.

The segmentation evaluation algorithm follows three steps. First each of the connected components returned by the method under evaluation is assessed based on whether it corresponds to a fraction, a whole, many or none of the ground truth defined text-parts. In the second step each text-part defined in the ground truth is classified as W*ell segmented*, *Broken*, *Merged*, *Broken & Merged* or *Lost* according to the results of the first step. The final step classifies atoms under the same categories based on the situation of their constituent text-parts.

4.1.1 Classification of Segmentation Components

Based on the above observations and making use of the dual representation of text parts (skeleton and area) in the ground truth, we introduce the *Minimal Coverage* and *Maximal Coverage* concepts. For a given text part defined in the ground truth, Minimal Coverage refers to the minimum set of pixels that a segmentation component should cover to be recognizable as this text part. This Minimal Coverage area is defined to be equivalent to the skeleton of the text part.

In a similar fashion for a given ground truth text part, Maximal Coverage refers to the maximum set of pixels that a segmentation component can cover to be recognizable as this text part. The Maximal Coverage area is algorithmically obtained by an expansion of the text part's area defined in the ground truth.

A critical aspect, as discussed before, is that ground truth information is not always accurate, both due to the human factor

[1] http://www.cvc.uab.es/~dimos/PerfEval_Schema.xsd

and due to the nature of the images (e.g. anti-aliased text as in Figure 4). To deal with this variability we introduced certain flexibility in how the Minimal and Maximal Coverage criteria work.

(a) **(b)** **(c)**

Figure 7. Cases where Minimal Coverage is (a) within or (b) outside the threshold.

The evaluation framework allows setting a threshold T_{min} on the percentage of skeleton pixels that a component should cover to pass the Minimal Coverage criterion. The default value for this threshold is 90% (Figure 7). Similarly, if a segmentation component comprises pixels that do not overlap with the area defined for the text-part in the ground truth, the framework allows setting a threshold T_{max} on the distance such component pixels are allowed to have from the area edge. If the distance is small, then the segmentation method is just producing a dilated version of the original text-part which would be acceptable (see Figure 3b), but if the distance is bigger then there is a protrusion that could alter the shape of the text-part and the component is rejected (see Figure 3c). T_{max} is set as a function of the thickness of the text-part capped to a maximum value and the default formula is $T_{max} = min(5, 0.5 \cdot G)$ in pixels, where G is the thickness of the text-part.

Each of the segmentation connected components is classified as *fraction*, *whole*, *multiple*, *fraction & multiple*, *mixed* or *background* according to the following heuristics.

If the segmentation component does not overlap with any of the text-part skeletons defined in the ground truth it is classified as *background*.

A segmentation component is classified as *whole* if it overlaps with a single text-part skeleton and the Maximal and Minimal Coverage criteria with the corresponding text-part are satisfied.

When a segmentation component overlaps with a single skeleton, the Maximal Coverage criterion is satisfied but the Minimal Coverage criterion is not satisfied, then the component is covering only a fraction of the corresponding text-part and it is labeled as *fraction*.

When a segmentation component overlaps with more than one skeleton the Minimal Coverage criterion is satisfied for each of the corresponding text-parts and the Maximal Coverage criterion is satisfied collectively for the combination of the corresponding text-parts, then the component is covering more than one text parts but without covering important parts of the background and it is therefore labeled as *multiple*.

When a segmentation component overlaps with more than one skeleton the Minimal Coverage criterion is not satisfied for at least one of the corresponding text-parts but the Maximal Coverage criterion is satisfied collectively for the combination of the corresponding text-parts, then the component is covering more than one text-parts without covering important parts of the background but does not cover all parts completely and it is therefore labeled as *fraction & multiple*.

In any other case, the connected component covers both non-text and text parts of the image and it is a bad segmentation result labeled as *mixed*.

Table 1. Classification of Connected Components.

Number of overlapping skeletons	Maximal Coverage Satisfied	Minimal Coverage Satisfied	Label
0	N/A	N/A	Background
1	Yes	Yes	Whole
1	Yes	No	Fraction
1	No	Yes	Mixed
1	No	No	Mixed
Many	Yes	Yes	Multiple
Many	Yes	No	Fraction & Multiple
Many	No	Yes	Mixed
Many	No	No	Mixed

The above heuristics are summarized in Table 1. In terms of segmentation, "background" and "whole" components are good results as the text and non-text areas are well separated, and "mixed" are bad results. The rest of the cases represent good separations of text from non-text areas, but indicate future problems in text extraction.

4.1.2 *Assessment of Text Parts*

Given the labeled segmentation components, the segmentation quality of each of the text-parts defined in the ground truth can be deducted. The text-parts are classified as *Well segmented*, *Broken*, *Merged*, *Broken & Merged* or *Lost* according to the following heuristics.

If there is no combination of *whole*, *fraction*, *multiple* and *fraction & multiple* connected components that cover at least T_{min} of the skeleton of the text-part in question, then the text-part is classified as *Lost*. In any other case, the classification is given by Table 2 in terms of the types of components that combined can cover the text-part in question.

Table 2. Classification of Text Part.

Whole	Multiple	Fraction	Fraction & Mult.	Label
Yes	N/A	N/A	N/A	Well Segmented
No	Yes	No	No	Merged
No	No	Yes	No	Broken
No	Yes	Yes	N/A	Broken & Merged
No	N/A	N/A	Yes	Broken & Merged
No	No	No	No	Lost

4.1.3 *Assessment of Atoms*

In the case of atoms that comprise a single text-part (classified as *Normal* or *Merged-by-design* in the ground truth), the classification of their text-part applies to the atoms as well. In the case of atoms that comprise more than one text-part (such as the *Broken-by-design* and *Broken & Merged-by-design* ones), an extra step is required to classify them into one of the previous classes.

The heuristics in Table 3 are used to do the final atom classification in terms of its constituent text-parts.

Table 3. Classification of Atoms.

If at least one text-part is *Lost*	Lost
Else if (at least one text-part is *Broken* **and** at least one text-part is *Merged*) **or** (at least one text-part is *Broken & Merged*)	Broken & Merged
Else if at least one text-part is *Broken*	Broken
Else if at least one text-part is *Merged*	Merged
Else	Well Segmented

4.1.4 Report of Final Results

The final results of the above classification process are given in terms of the percentages of ground truth atoms classified under each of the five categories defined above. This kind of results carries much more informative value than pixel misclassification rates as they directly relate to the ultimate goal of character recognition. At the same time they allow for a qualitative evaluation, while they still offer a way to quantify the overall segmentation quality.

Importantly, a side-effect of this evaluation method is that generic colour segmentation methods can also be assessed through the same framework, since the raw material for the evaluation is just a list of components. There will be no false negatives in this case as all components would be considered. That should be taken into account, but other than that the evaluation would work in exactly the same way.

4.2 Assessment of text extraction performance at the atom level

Following the evaluation in terms of segmentation, the framework can assess the capability of the text extraction method to correctly label and group text-parts into atoms. For the segmentation evaluation we were looking at whether each of the atoms defined in the ground truth was present in the segmentation results. Now we need to assess each of the atoms returned by the method in terms of whether they correctly match any of the ground truth defined atoms or not.

Each of the atoms returned by the text extraction method can be either *Correct* (true positive) if it matches any of the atoms defined in the ground truth, or *Wrong* (false positive) if it doesn't.

Building upon the connected component classification described in section 4.1.1, we can define the condition for an extracted atom to match a ground truth atom as follows:

- All connected components grouped under the extracted atom are *whole* connected components, and
- All connected components of the extracted atom correspond to text-parts of the same ground truth atom, and
- The ground truth atom does not have more text-parts than the extracted atom has components.

Assuming N_{GT} the number of atoms in the ground truth, N_{RET} the number of atoms returned by the text extraction method and C the number of them which are *Correct*, we can define the following metrics for the text extraction method, where P is precision, R is recall and F is the F_1-score (giving equal importance to precision and recall).

$$P = \frac{C}{N_{RET}} \quad (1)$$

$$R = \frac{C}{N_{GT}} \quad (2)$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \quad (3)$$

4.3 Assessment of text extraction performance at the word and text line level

If the text extraction method is able to return words and text-lines as groups of atoms, then correctly / incorrectly extracted words and text-lines can be decided in the same fashion as atoms before. Namely, if all atoms comprising a returned word are classified as *Correct*, they all pertain to the same ground truth word, and the ground truth word in question contains no more atoms than the extracted word, then the extracted word is *Correct*. This definition can similarly be extended for returned text-lines. The definitions of Precision, Recall and F-score would be the same.

On the other hand, if the text method is only able to return bounding rectangles, then comparison can be made at the bounding rectangle level, following any of the performance evaluation techniques developed for text localisation algorithms (e.g. the one Wolf [8]).

5. CASE STUDY

Following the definition of the complete text extraction performance evaluation framework, we demonstrate below its potential use in assessing a text extraction method. This is not meant to be a comprehensive evaluation of a state of the art text extraction method, but merely a case study to illustrate the use of the framework. In this instance we are using the method described in [2] as the candidate for evaluation.

Table 4. Summary of Atoms in the Ground-Truth.

Image ID	Normal	Broken by Design	Merged by Design	Total
1	20	2	2	24
2	27	1	0	28
3	14	1	0	15
4	11	2	0	13
5	20	0	0	20
6	41	5	0	46
7	30	0	0	30
8	10	0	0	10
9	20	0	0	20
10	32	0	0	32
11	20	0	0	20
12	19	1	0	20
13	23	1	0	24
TOTAL	**287**	**13**	**2**	**302**

A comprehensive evaluation is difficult here for a number of reasons, most importantly the fact that there are still no representative and statistically adequate datasets to base such an analysis on. We are in the process of ground truthing and releasing to the community such datasets (see also next section), and expect that some will be available by the time of the conference. For the case study we have ground-truthed a small part (13 images) of the free ICDAR dataset used for the "ICDAR 2003 Robust Reading competition" [7], [14]. A summary of the dataset in terms of atoms, as compiled automatically from the ground truth information is shown in Table 4.

5.1 Assessment of segmentation performance

First, we are assessing the segmentation capabilities of the method as explained in section 4.1. The objective here is to assess the text extraction method in terms of its capacity to separate text from non-text areas at the pixel level (lost atoms vs the rest), while maintaining the integrity of the text parts by not over-segmenting or under-segmenting the image (tendencies to produce *Broken*, *Merged* or *Broken & merged* atoms). Table 5 demonstrates the type or results attainable in terms of segmentation.

Table 5. Segmentation Performance.

Image ID	Well Segm.	Broken	Merged	B & M	Lost
1	8.3%	33.3%	0.0%	12.5%	45.9%
2	63.0%	0.0%	0.0%	0.0%	37.0%
3	78.6%	0.0%	0.0%	0.0%	21.4%
4	100.0%	0.0%	0.0%	0.0%	0.0%
5	90.0%	0.0%	0.0%	0.0%	10.0%
6	92.7%	0.0%	0.0%	0.0%	7.3%
7	43.3%	3.3%	0.0%	0.0%	53.3%
8	90.0%	0.0%	0.0%	0.0%	10.0%
9	95.0%	5.0%	0.0%	0.0%	0.0%
10	96.9%	0.0%	0.0%	0.0%	3.1%
11	80.0%	0.0%	0.0%	0.0%	20.0%
12	100.0%	0.0%	0.0%	0.0%	0.0%
13	75.0%	0.0%	0.0%	0.0%	25.0%
TOTAL	**76.0%**	**3.4%**	**0.0%**	**1.0%**	**19.5%**

5.2 Assessment of Atom Extraction

Following the segmentation analysis, we can now study the behavior of the method under testing at the atom extraction level, namely its ability to interpret the segmentation results and re-combine the constituent parts of text correctly. Two different types of results can be obtained here, one is the performance per image, and another is the performance over the whole dataset in terms of the category of the atoms (as labeled in the ground-truth).

The first summary is revealing of the relative importance of the segmentation and the extraction steps in each case. For example in the case of image #1, the low text extraction score is easily attributed to the segmentation results, but in the case of image #11 segmentation cannot explain the failure. It turns out that the texture pattern in the image is causing the text extraction part of the algorithm to produce a lot of false text-lines (see Figure 8).

The second summary is useful to identify systematic failures for a specific atom category (*Normal*, *Broken-by-design*, *Merged-by-design*), i.e. assess the efficiency of dealing with inherent difficulties due to the way atoms are designed. The algorithm under testing here is pretty simplistic in its atom extraction process as it assumes that atoms comprise a single text-part and it is unable to put together characters such as "i" and "j". This is easily observed here as precision and recall is zero for the *Broken-by-design* atoms (although the dataset is too small to allow for any statistically relevant conclusion to be made).

Table 6. Atom Extraction Results per Image.

Image ID	Precision	Recall	F-score	*Segmentation Performance*
1	0.01	0.08	0.02	*8.3%*
2	0.3	0.46	0.36	*63.0%*
3	0.17	0.73	0.27	*78.6%*
4	0.28	0.77	0.41	*100.0%*
5	0.33	0.9	0.49	*90.0%*
6	0.58	0.74	0.65	*92.7%*
7	0.13	0.43	0.2	*43.3%*
8	0.1	0.9	0.19	*90.0%*
9	0.12	0.95	0.21	*95.0%*
10	0.36	0.97	0.53	*96.9%*
11	0.05	0.8	0.09	*80.0%*
12	0.35	0.95	0.51	*100.0%*
13	0.02	0.54	0.03	*75.0%*
TOTAL	**0.1**	**0.7**	**0.18**	

Table 7. Atom Extraction Results per Atom Category.

Atom Category	Precision	Recall	F-score
Normal	0.1	0.7	0.18
Broken by Design	0	0	0
Merged by Design	0	0	0

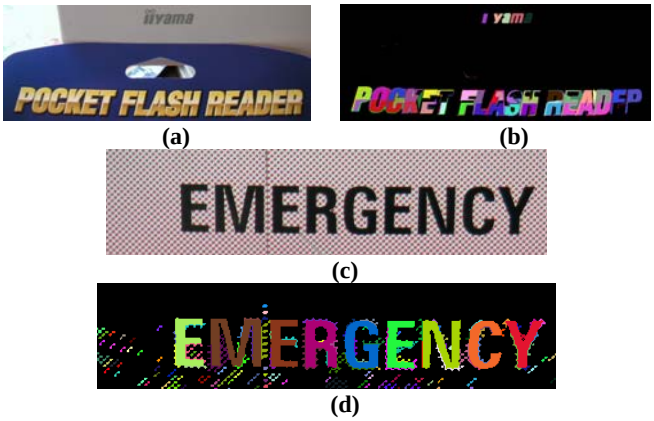

Figure 8. (a) Detail of image #1 of the dataset and (b) result of the text extraction method. (c) Detail of image #11 of the dataset and (d) result of the text extraction method.

Assessment at levels higher than atoms is possible as described in section 4.3, including assessment at the level of bounding polygons when this is applicable. For the Word and Text Line level results similar to the tables above are attainable by the framework.

6. DISCUSSION

As can be easily appreciated, the dataset used here is not by any stretch of imagination complete in the sense that it does not cover all different situations imagined. For example there are not many *Merged-by-design* or *Broken-by-design* atoms. Two observations should be made at this point. First, that in order for this performance evaluation framework to become a useful tool for the community, a number of distinctly different datasets (representative of the real world and statistically adequate) are required, that cover different application domains. We are in the process of ground-truthing and releasing such datasets to the community, including the ICDAR'03 Robust Reading dataset, scanned posters, and Web images. These are expected to be submitted to the TC11 Web site as soon as they are ready.

Second, it is useful to observe that given a dataset which is representative of a specific application, useful information can be deducted from the summary of the ground-truth. For example real-scene images (used in the scenario above) do not suffer from many *Broken-by-design* or *Merged-by-design* atoms, while Web images contain a considerable amount of them. This is useful on its own right, to define the open issues and do research planning. As stated before, we are in the process of ground truthing and releasing to the community datasets covering many different application domains. We hope that the final result will be useful for the community in terms of highlighting the open issues in the different application domains.

The framework as it stands is quite complete. Assessment can be done at different levels, and many real-life situations can be specified in the ground truth and used for qualitative assessment. There are nevertheless a number of special cases that are still not encoded in the ground truth or dealt with in the evaluation. Take for example the "ebay" image of Figure 1b. A human observer easily identifies four individually produced characters in the word "ebay" that *overlap*. Should we penalise a text extraction algorithm for assigning the common part between "b" and "a" to one of the two characters and not to both? Does it really belong to any of the two characters or should it be included twice in the ground truth? Cases like these exist, but are rare and rather application-domain specific. It is a question of usage whether increasing the complexity of the framework versus including all possible scenarios is a good trade-off or not.

7. CONCLUSIONS

In this paper we present a ground truth specification and performance evaluation framework for assessing the performance of text extraction methods on complex colour images. The framework allows fast automatic comparisons between methods. It is able to report qualitative and quantitative results at different levels of the text extraction process, and reveal where problems lie in the text extraction process. A case study based on the ICDAR'03 Robust Reading was presented to demonstrate the usage of the framework in a real scenario.

8. ACKNOWLEDGMENTS

This work was partially supported by the Spanish projects TIN2008-04998, TIN2009-14633-C03-03, CONSOLIDER INGENIO CSD2007-00018 and the fellowships RYC-2009-05031 and 2009FI-B00020. The authors would also like to thank B. Gatos and A. Antonacopoulos for providing useful insight.

9. REFERENCES

[1] T. Retornaz and B. Marcotegui, "Scene text localization based on the ultimate opening", Proceedings of the 8th International Symposium on Mathematical Morphology, Rio de Janeiro, Brazil, Oct. 10 –13, 2007, Vol. 1, pp. 177–188.

[2] A. Clavelli and D. Karatzas, "Text Segmentation in Colour Posters from the Spanish Civil War Era", Proceedings of the 10th International Conference on Document Analysis and Recognition, IEEE CPS, pp.181–185, 2009

[3] S.J. Perantonis, B. Gatos, V. Maragos, V. Karkaletsis and G. Petasis, "Text Area Identification in Web Images", in "Methods and Applications of Artificial Intelligence", LNCS 3025, pp.82–92, 2004

[4] D. Lopresti and J. Zhou, "Locating and Recognizing Text in WWW Images," Information Retrieval, vol. 2, 2000, pp. 177-206

[5] D. Karatzas, A. Antonacopoulos, "Colour Text Segmentation in Web Images Based on Human Perception", Image and Vision Computing, 25(5), pp. 564-577, 2007

[6] Kwang In Kim, Keechul Jung, and Jin Hyung Kim, "Texture-Based Approach for Text Detection in Images Using Support Vector Machines and Continuously Adaptive Mean Shift Algorithm", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No, 12, 2003

[7] S.M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, H. Miyao, J. Zhu, W. Ou, C. Wolf, J.M. Jolion, L. Todoran, M. Worring, and X. Lin, "ICDAR 2003 robust reading competitions:entries, results, and future directions", International Journal on Document Analysis and Recognition Vol. 7, no. 2-3, pp. 105–122, 2005

[8] C. Wolf, J.-M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms", International Journal of Document Analysis, 8(4), pp.280–296, 2006

[9] A. Antonacopoulos, D. Karatzas and D. Bridson, "Ground truth for layout analysis performance evaluation", In Proceedings 7th IAPR Document Analysis Workshop, 2006

[10] J. Liang, I. Phillips, R. Haralick, "Performance evaluation of document layout analysis algorithms on the UW data set", In: Proceedings of SPIE, Document Recognition IV, pp 149–160, 1997

[11] F. Einsele, R. Ingold, J. Hennebert, "A Language-Independent, Open-Vocabulary System Based on HMMs for Recognition of Ultra Low Resolution Words", Journal of Universal Computer Science, vol. 14, no. 18, pp. 2982-2997, 2008

[12] M.A. Moll, H.S. Baird and C. An, "Truthing for Pixel-Accurate Segmentation", Proceedings of the 8th International Workshop on Document Analysis Systems, IEEE CPS, pp.379–385

[13] K. Ntirogiannis, B. Gatos and I. Pratikakis, "An Objective Evaluation Methodology for Document Image Binarization Techniques", Proceedings of the 8th International Workshop on Document Analysis Systems, IEEE CPS, pp.217–224

[14] http://algoval.essex.ac.uk/icdar/Datasets.html