Document perceptual quality ground truth creation

Vincent Rabeux
LaBRI
Bordeaux, France
rabeux@labri.fr

Nicholas Journet
LaBRI
Bordeaux, France
journet@labri.fr

Anne Vialard
LaBRI
Bordeaux, France
vialard@labri.fr

Jean-Philippe Domenger
LaBRI
Bordeaux, France
domenger@labri.fr

***Abstract*—This article focuses on a new method for document perceptual quality ground truth creation. This type of ground truth gives a quality related score to each image in a dataset. This is useful for performance evaluation of algorithms that measure the quality of images. The quality of a document image is related to the amount of its degradations.**

To our knowledge, a methodology to create this kind of database, specific to document, does not exists. Every known method proposes empirical and subjective protocols. Moreover, the creation of these ground truths takes a very long time.

In this article, we present a new methodology to create this kind of ground truth. This methodology has two main advantages : it minimizes, both, the time spent to create such ground truth and the subjectivity in respect to traditional methods. The time and subjectivity are lowered by using a binary search insertion sort ($log2(N)$ comparisons maximum). A user only has to select within two images the one that is the most degraded (according to a quality criteria). Moreover, the tool presented in this article is implemented using web services allowing the creation of ground truths in a collaborative way.

***Keywords*-document; ground truth; image; quality; perceptual;**

I. Introduction

In order to evaluate the quality of document images, algorithms proposed in the state of the art are based on objective and quantifiable criteria [4] : structural similarity index (gap to a reference image), contours analysis (without reference image). The authors of [4] are interested in the evaluation of these algorithms that apply to a wide variety of images (2D images, animated images, 3D images). In general, the ground truth is acquired by showing a set of images to a set of users. Users rate the images on a predefined quality scale. The quality of an image is, at last, the mean opinion score (MOS). One can then evaluate an algorithm that produce a measure of visual quality by analyzing the correlation between the algorithm results and the ground truth created by the users.

This methodology has several disadvantages. First the creation of a complete dataset is a very tedious work. Second, the users need to have a global knowledge of the images that need to be rated. This freezes the ground truth in time since the addition of new images can invalidate the previous judgements (the ground truth is hard to maintain).

In [5] several users have to evaluate the quality of each distorted version of a reference image (68 images). The originality of the proposed methodology lies in the fact that scores given by users are not absolute but relative. Indeed, two distorted images are compared to the reference image and the user has to choose the image that is the closest to the original. The *swiss competition principle* is then used to limit the number of images pairs to be compared. At the end of the process, each image is rated with a score between 0 and 9. If this methodology speeds up the time spent to create these ground truths, they are still unmaintainable (how to manage the insertion of a new image ?).

In this article, we present a ground truth creation methodology well suited to document images, in witch users sort a set of images using a perceptual and visual quality criteria. This methodology allows a fast creation of a ground truth that can be maintained. At last the implementation of this methodology relies on web services allowing the creation of the ground truth in a collaborative way. We first detail the methodology and how a ground truth can be created. We then propose to study some use cases in order to verify that the proposed methodology is accurate.

II. Perceptual ground truth creation methodology

In order to create ground truths that can be easily maintained and enriched in time, we decided not to assign an absolute score to images but to build the ground truth by using successive image comparisons. Moreover, this methodology does not need a reference image like in [5]. Images are compared by the users by answering a question on a specific quality criteria. We propose a unique technical environment (iPad) in order to minimize visual and perceptual difference due to the screen size, resolutions, ...

The proposed procedure then creates a list of images that are sorted following a perceptual quality criteria. The over-all process follows the *binary search insertion sort* that consist in using a dichotomic search to determine where to insert a new image. The comparison between two images is made by the user by selecting the image that answers the best to a question (*ie.* which image contains the most bleed through ?). This algorithm takes advantage of the fact that the sub list is already sorted and yields to an efficient way to sort images in terms of number of comparisons. Indeed, a new image is inserted after $log2(N)$ comparisons maximum.

The implementation of this algorithm is based on web services. This allows the ground truth to be created by several users at the same time and in a collaborative way. This raises the problem of inter-rater variability. In order to circumvent this problem, we propose to clone the set of images to sort into n lists. Each image has to be sorted in every list by different users. The final index of an image is its mean index of all indexes where it has been inserted.

III. Experiments and use cases

To evaluate this methodology, we created two datasets of semi-synthetic images with their corresponding ground truth (the degradations models parameters). The first one addresses the JPEG2000 compression algorithm and the second one contains bleed through images generated using the degradation model presented in [3]. The JPEG2000 dataset contains 24 images of documents compressed on 8 different levels. The bleed through data set contains a total of a 100 images with 4 different levels of intensity (an example of such images can be seen on figure 1).

Figure 1. First row : Different levels of bleed through intensity (from left to right : low, medium, high). Second row : different levels of JPEG compressions

Several users were asked to compare either the degree of the compression (ground truth 1), or the bleed-through intensity (ground truth 2) of images contained in the two datasets. We then used the $kappa$ statistical test in order to measure the agreement between observed qualitative judgments. The kappa test results in the sum of two components : the agreement expected by chance and the actual agreement (equation 1).

$$K = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

With :

- P_o : the actual agreement.
- P_e : the agreement expected by chance.

The kappa coefficient K is between -1 and 1. The more its value is close to 1, the more users agree with the real ground truth. A good agreement is estimated when the Kappa coefficient is above 0.60 [2]. At last the kappa coefficient is here weighted (squared kappa - [1]) so that a small disagreement (just a few indexes) is less important that a big disagreement.

List	1	2	3	4	Merged
Kappa (JPEG2000 dataset)	0.78	0.82	0.86	0.68	0.88
Kappa (bleed-through dataset)	0.88	0.93	0.85	0.75	0.93

Table I
Every user agrees with the real ground truth (> 0.60). Kappa values that are higher that 0.80 can be considered as excellent agreements (Cohen et al., 1960).

In the results presented in Table I, each user's list agrees with the real ground truth ($kappa > 0.60$). This fact means that the way we sort images in a list is a good way to create perceptual ground truths. Moreover, we can see that the merged list (the list obtained with the mean index of each images) has an even better kappa coefficient. This means that the use of several lists minimizes the errors made by some users.

IV. Conclusion and perspectives

In this paper a new, fast and collaborative methodology to create ground truth related to human quality perception is presented. Tests on synthetic data shows that it is accurate enough to be used in a ground truth creation campaign. The main perspective of this work is to improve the list merge step by detecting outliers.

References

[1] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[2] J. Cohen et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[3] Reza Moghaddam and Mohamed Cheriet. Low quality document image modeling and enhancement. *International Journal on Document Analysis and Recognition*, 11:183–201, 2009.

[4] Anush Moorthy and Alan Bovik. Visual quality assessment algorithms: what does the future hold? *Multimedia Tools and Applications*, 51:675–696, 2011.

[5] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. Tid2008 - a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10:30–45, 2009.