Statistic Metrics for Evaluation of Binary Classifiers without Ground-Truth

Maksym Fedorchuk
NTUU "Igor Sikorsky Kyiv Polytechnic Institute" –
Faculty of Biomedical Engineering,
Kyiv, Ukraine
m.fedorchuk-2017@kpi.ua

Bart Lamiroy
Université de Lorraine – LORIA (UMR 7503)
Nancy, France
Bart.Lamiroy@loria.fr

***Abstract* – In this paper, are presented a number of statistically grounded performance evaluation metrics capable of evaluating binary classifiers in absence of annotated Ground Truth. These metrics are generic and can be applied to any type of classifier but are experimentally validated on binarization algorithms. The statistically grounded metrics were applied and compared with metrics based on annotated data. This approach has statistically significant better than random results in classifiers selection, and our evaluation metrics requiring no Ground Truth have high correlation with traditional metrics. The experiments were conducted on the images from the DIBCO binarization contests between 2009 and 2013.**

***Keywords* – statistical evaluation; performance evaluation; image binarization; segmentation; document image analysis.**

I. Introduction

In machine learning and statistics, classification is the problem of identifying to which of a set of categories a new element belongs. In many cases, this is a fully automated task and there are many different classification systems, developed for specific tasks and data types. Usually, there exist a huge number of possible classifiers for a given task, and it is important to be able to determine the ones most suitable for the problem at hand. In this paper, we will try to determine if standard performance metrics can be extended to situations where no reference benchmark data (*Ground Truth*) are available.

Indeed, the traditional approach to assessing the validity and performance of classifiers generally consists of 3 phases:

- assemble a representative collection of reference data;
- use human annotators to create a set of reference interpretations for the data collection (*Ground Truth*);
- run the classifiers to be evaluated on the reference data and measuring their discrepancy with the expected *Ground Truth*.

While this paradigm is well understood, and has been largely adopted to assess and measure the quality of state-of-the-art classifiers, it has a number of limitations and drawbacks [1], one of which is that it assumes *Ground Truth* is void of errors.

The work presented in this paper revisits the concepts presented in [1] by investigating the idea of comparing and evaluating classifiers if no *Ground Truth* is available and focus on binary classifiers (*i.e.* classifiers that arrange elements in only two categories). We will be reviewing some of the commonly used metrics for measuring discrepancy: F-Measure, Peak Signal-to-Noise Ratio, Normalized Cross Correlation and Negative Rate Metric, and establish whether they can be approximated by statistical counterparts. We experimentally validate our approach by applying these proposed metrics to a specific type of binary classifier: image binarization. The reader should be aware that findings of this paper do generalize to any kind of binary classifier, however.

Image binarization is a preprocessing step, often necessary for document analysis, medical image analysis, pattern recognition, computer vision and any other, content-from-image extraction systems. It converts the image in a bi-level form for further image processing. Depending on the type of images or the type of subsequent needs down the image treatment pipelining it can be a daunting task to decide which binarization method to apply, even using adaptive systems with multiple algorithms of binarization. Unsuitable binarization can be in many cases the reason to fail forthcoming processing steps or reduce their performance.

For the experiments, we will be using state-of-the-art binarization algorithms in the light of the previously described evaluation protocol. The simplest way of estimating which one is more is suitable for a given type of image is to take an image with known *Ground Truth* (GT) and evaluate the quality by standard comparison metrics. However, in cases where we have no access to suitable images or we have not precise enough Ground Truth [1], [2], [3] it is necessary to use alternative tools. This paper develops the idea of using statistical tools for evaluation by calculating precision and recall [4], [5] without GT and using the consensus coming from multiple classification systems as a reference [6].

The rest of this paper is organized as follows: first, we provide an overview of the experimental framework, and the collection of binarization algorithms that were used, followed by the analysis of standard existing performance metrics for which we suggest statistical counterparts in Section IV. Section V provides experimental validation and a conclusion is given in Section VI.

II. Tested Binarization Algorithms

It is common to categorize binarization algorithms in *global* and *local* methods. The general approach for every

binarization system is similar: if a pixel (m, n) in the input image has a higher gray level value than a given threshold, then this pixel labeled as background, otherwise, it is labeled as foreground. Individual binarization approaches differ in how the threshold is computed: global algorithms calculate one threshold for the entire image, while local algorithms calculate different threshold values for each pixel of the image, depending on their surrounding region. Global binarization is faster and simpler to implement than local algorithms.

Multiple global binarization algorithms exist in the literature, based on various classification procedures: histogram operations; clustering; entropy analysis and Gaussian distributions. But according to the results of their performance in DIBCO images dataset we decided to retain only one: global Otsu`s binarization method.

Locally adaptive binarization methods are able to compute a threshold for each separate pixel using the information contained in neighborhood pixels and therefore usually show better performance than global ones.

This paper is based one global and nine locally adaptive algorithms:

1) Otsu's global binarization method [8];
2) Local Otsu's method;
3) Brensen's method [9];
4) Niblack's method [10];
5) Breadly's method [11];
6) Method of local medium value;
7) Modified Gato's method;
8) Wolf's method [12];
9) Kittler's method [13];
10) Sauvola's method [14];

The selected algorithms have significant reported performance differences. All of them can show high enough quality for some types of images according to the DIBCO evaluation campaigns. Local Otsu's method and Gato's method were slightly modified with respect to their published versions, as will be explained below.

Local Otsu's method is based on a global threshold binarization method described by N. Otsu. The idea of selecting an automatic threshold level according to this method was based on the analysis of the image graylevel histogram. Traditionally, the threshold is selected by maximizing the measure of separability between the classes in graylevel. Here, a modified version of the method was used: a local threshold level was chosen for every pixel by applying Otsu's method over a local square window. The optimal size found to be good a choice was 100 by 100 pixels.

Gato's method [15] includes several distinct steps. It is a pre-processing procedure using a low-pass Wiener filter; estimation of foreground regions; background surface calculation by interpolating neighboring background intensities and thresholding by combining the calculated background surface with the original image while incorporating image up-sampling.

Gato's method was modified by adding histogram equalization and median filtering after using a low-pass Wiener filter. Optimal size for low-pass and Wiener filter was 5 by 5 and 17 by 17 pixels.

All the chosen algorithms depend on input parameters and their performance is sometimes sensitive to subtle changes. We applied consistent and near-optimal parameters for all experiments, either by applying the recommended published parameters, either experimentally determined parameters. It should be clear to the reader that the scope of this paper concerns Ground Truth-less performance metrics, and not binarization. Therefore, whether the choice of parameters is actually optimal or not is of no incidence to the conclusions drawn from our study on the various metrics developed in Section IV.

III. Input Data

In the experiments reported in Section V, we use all 56 images from the Digital Image Binarization Contest (DIBCO) editions between 2009 and 2013. DIBCO is organized in the context of the International Conference of Document Analysis and Recognition (ICDAR). The general objective of the contest is to identify advances in document image binarization by applying evaluation performance measures. All editions focus on the evaluation of document image binarization methods using a variety of scanned machine-printed and handwritten documents for which the organizers created the *binary Ground Truth* images using a semi-automatic procedure [7]. Binarization methods competing in DIBCO are compared to each other in function of 4 metrics. These metrics are the F-Measure, PSNR, NCC and NRM, and consist of different means of measuring the discrepancy of various binarization outputs (resulting from the competing methods) with the established *Ground Truth*.

One of the motivations of this paper is that this approach has been challenged with as main argument that the *Ground Truth* cannot be considered unique and therefore that subsequent performance evaluation is biased [1], [3] .

IV. Evaluation Metrics

Given the legitimate objections to Ground Truth-based evaluation expressed in [1], [2], [3] we explore the idea of using performance metrics that can be used in absence of Ground Truth as has already been experimented in [4]. The main idea behind the approach is to replace the standard Ground Truth with a consensus metric resulting from the collection of compared methods. This section reviews the conventional performance metrics and reformulates them in the context of this idea of consensus metric. In Section V will then measure and experimentally establish their validity.

A. F-Measure

F-measure combines precision and recall by calculating their harmonic mean:

$$FM = \frac{2 * Precision * Recall}{Prcision + Recall} \quad (1)$$

Calculation of Precision and Recall is based on the relation between true and false determined elements. Precision is the value of retrieved elements that are relevant and Recall is the value of relevant elements that are retrieved.

$$Precesion = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

TP, FP, FN denote the True Positive, False Positive and False Negative values, respectively.

B. PSNR

In image analysis, Peak Signal-to-Noise Ratio (PSNR) is the maximum value between the power of a signal and corrupting noise. In our case, this ratio measures how close an estimator image is to the estimated image. It is expressed in terms of the logarithmic decibel scale:

$$PSNR = 10log(\frac{C^2}{MSE}) \quad (4)$$

Where C^2 is a maximum possible value of separate pixels (difference between foreground and background), and the mean squared error (MSE) described by following equation:

$$MSE = \frac{\sum_{x=1}^{M} \sum_{y=1}^{N} (A(x,y) - B(x,y))^2)}{MN} \quad (5)$$

The higher are value of PSNR, the higher is the similarity of the two images.

C. NCC

Normalized Cross Correlation is often used for comparing multidimensional arrays and is defined by the following equation:

$$NCC = \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} (A(m,n) - \bar{A})(B(m,n) - \bar{B})}{\sqrt{\sum_{m=1}^{M} \sum_{n=1}^{N} (A(m,n) - \bar{A})^2 \sum_{m=1}^{M} \sum_{n=1}^{N} (B(m,n) - \bar{B})^2}} \quad (6)$$

Where $\bar{A}$ is the mean value of one array and $\bar{B}$ is the mean of another, and M and N are the dimensions of the arrays.

A higher NCC indicates better matching of arrays.

D. NRM

The Negative Rate Metric is a numerical equivalent of the relation between misclassified elements and all other elements in the class. It is the average value of two negative rates: false negative rate NR_{FN} and false positive rate NR_{FP}:

$$NRM = \frac{NR_{FN} + NR_{FP}}{2} \quad (7)$$

$$NR_{FN} = \frac{FN}{TP + FN} \quad (8)$$

$$NR_{FP} = \frac{FP}{TN + FP} \quad (9)$$

A higher NRM indicates a worse mismatch between two classifiers. Section V provides the experimental validation of their relevance.

E. Pseudo F-Measure

We are using the concept of Pseudo F-Measure introduced in [1]. The idea developed in this paper is that the statistical equivalent of ground truth is an array of probabilities that the documents (separate pixels in image binarization) δ_i belongs to the foreground cluster Δ^+. Under this hypothesis these probabilities are:

$$P(\delta_i) = \frac{1}{s} \sum_{k=1..s} S_k(\delta_i) \quad (10)$$

Where $S_k(\delta_i)$ represents the classification result of document δ_i by classifier S_n, and s the number of classifiers.

Given the hypothesis of equivalent distribution of all documents δ_i in their set Δ^*, the authors state that *Precision* "*is the probability that a random document retrieved by a query is relevant*", and define $Pr(S_k)$ as:

$$Pr(S_k) = \frac{\sum_{i=1..d} P(\delta_i) S_k(\delta_i)}{\sum_{i=1..d} S_k(\delta_i)} \quad (11)$$

Where d is the total number of elements (in our case, pixels in the image) classified by classifier S_k as belonging to one of classes (foreground or background).

Similarly, *Recall* was defined as "*the probability for a random relevant document to be retrieved by the query*" and described by the formula:

$$Rc(S_k) = \frac{\sum_{i=1..d} P(\delta_i) S_k(\delta_i)}{\sum_{i=1..d} P(\delta_i)} \quad (12)$$

In these cases, relevancy has been replaced by $P(\delta_i)$.

These two described statistically-grounded metrics can be combined into *Pseud F-Measure* in a similar way as for the ordinary F-measure by computing the harmonic mean between Precision and Recall (1).

Besides this extension of [4], we have added three more statistical metrics for classifier evaluation: Pseudo Negative Rate Metric (Pseudo NRM), Pseudo Normalized Cross Correlation (Pseudo NCC) and Pseudo Peak Signal-Noise Rate (Pseudo PSNR).

F. Pseudo NRM

The statistically grounded alternative to Negative Rate Metric should be defined in function of the statistical equivalents of False Negative, False Positive, True Positive and True Negative values. In this case, relevancy also was replaced by $P(\delta_i)$ which had been described above (10).

According to our assumptions, and in accordance with of [3], the value of the Pseudo *True Positive* for a given classifier S_k can be expressed as the dot product of the array of $P(\delta_i)$ and the array given by the classification results $S_k(\delta_i)$. In the same way, the value of the Pseudo *False Negative* determined elements is the result of the dot product of $P(\delta_i)$ with the array of inverted binary result given by the classifiers S_k.

These two values allow for computing False Negative rates:

$$Pseudo\ NR_{FN} =$$

$$\frac{\sum_{i=1..d} P(\delta_i)\,\overline{S_k(\delta_i)}}{\sum_{i=1..d} P(\delta_i)\,\overline{S_k(\delta_i)} + \sum_{i=1..d} P(\delta_i)\,S_k(\delta_i)} = \frac{\sum_{i=1..d} P(\delta_i)\,\overline{S_k(\delta_i)}}{\sum_{i=1..d} P(\delta_i)} = 1 - \frac{\sum_{i=1..d} P(\delta_i)\,S_k(\delta_i)}{\sum_{i=1..d} P(\delta_i)} \quad (13)$$

One can observe that NR_{FN} actually is (1-Recall) and that this translates into Pseudo NR_{NF} being (1-Rc).

We can express the Pseudo Negative Rate of False Positive elements similarly:

$$Pseudo\ NR_{FP} = \frac{\sum_{i=1..d} (1 - P(\delta_i))\,S_k(\delta_i)}{\sum_{i=1..d} (1 - P(\delta_i))S_k(\delta_i) + \sum_{i=1..d} (1 - P(\delta_i))\,\overline{S_k(\delta_i)}} \quad (14)$$

We can simplify this equation to

$$Pseudo\ NR_{FP} = \frac{\sum_{i=1..d} (1 - P(\delta_i))\,S_k(\delta_i)}{d - \sum_{i=1..d} P(\delta_i)} \quad (15)$$

The final equation for Pseudo NRM is the same as for ordinary NRM (7) and is described as the average of the negative rate of false positive and false negative values:

$$Pseudo\ NRM = \frac{Ps\_NR_{FN} + Ps\_NR_{FP}}{2} \quad (16)$$

In contrast to F-Measure and PSNR, the lower the value for this metric, the better the classifier.

G. *Pseudo NCC*

Pseudo Normalized Cross Correlation expresses the level of normalized correlation between the probability that the elements δ_i belongs to the foreground cluster Δ^+ given the majority voting $P(\delta)$ and the result given by classifier S_k. Its expression is:

$$Pseudo\ NCC = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}(S_k(m,n) - \overline{S_k})(P_\delta(m,n) - \overline{P_\delta})}{\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{N}(S_k(m,n) - \overline{S_k})^2 \sum_{m=1}^{M}\sum_{n=1}^{N}(P_\delta(m,n) - \overline{P_\delta})^2}} \quad (16)$$

The higher this value, the better both arrays correlate with each other.

H. *Pseudo PSNR*

PSNR is a measure expressing how close one image is to another, *Pseudo PSNR* measures how close the result of a classifier is with respect to an array of probabilities, based on majority voting $P(\delta)$ described above (10). It is defined by the following equation:

$$Pseudo\ PSNR = 10log(\frac{C^2}{MSE}) \quad (17)$$

Where *MSE* is the mean squared error given by the average of the squares between evaluated array S_k and $P(\delta)$:

$$MSE = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}(S_k(m,n) - P_\delta(m,n))^2)}{MN} \quad (18)$$

The higher the value of Pseudo PSNR, the higher the similarity of the two arrays.

The assumption is that the performance of all of these described pseudo-metrics will give acceptable results for evaluation of classification systems without using ground truth or any annotated data at all. The next section describes the experimental protocol and conducted experiments that establish this.

V. EXPERIMENTS

In order to compare above described evaluation metrics we use the correlation coefficient $\boldsymbol{r}$ between each metric and its statistically-grounded pseudo-metric for each image in every dataset. We computed their correlation coefficient using the equation of normalized cross correlation, described in (6). For every picture in every dataset, we applied all the binarization systems, mentioned in paragraph II and applied each of the eight metrics described in paragraph IV to all obtained results. After that, was computed the normalized cross-correlation on each couple of metric and its corresponding pseudo-metric. The resulting correlation values for each dataset are shown in Fig. 1.

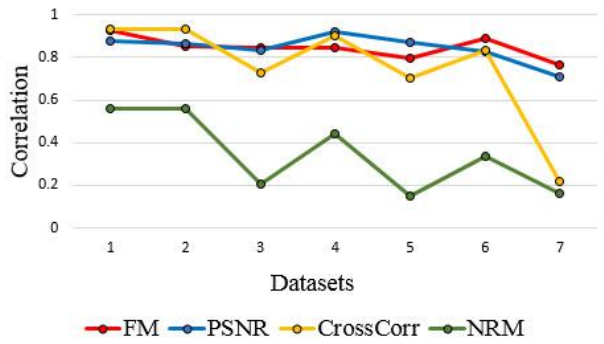

Fig. 1. Correlation between GT-based and statistical metrics for the all DIBCO datasets

The results obtained from all the datasets are also shown in Table 1. Rows represent datasets of images from DIBCO competition editions of 2009, 2011, 2012 and 2013. All the data consist of four hand-written (HW) and three printed (Pr.) sets of images.

TABLE I. AVERAGE CORRELATION COEFFICIENT OF EVERY METRIC AND PSEUDO-METRIC

	Average correlation coefficient			
DIBCO	***FM & Pseudo FM***	***PSNR & Pseudo PSNR***	***CrossCorr & Pseudo CrossCorr***	***NRM & Pseudo NRM***
09'Pr.	**0.93**	0.88	**0.93**	0.56
11'HW.	0.85	0.86	**0.93**	0.56
11'Pr.	**0.85**	0.83	0.72	0.21
12'HW	0.85	**0.92**	0.90	0.44
13'HW	0.79	**0.87**	0.70	0.15
13'Pr.	**0.89**	0.83	0.84	0.34

09'HW	**0.76**	0.71	0.22	0.16
Average	0.845	**0.856**	0.783	0.373
St. deviation	**0.051**	0.060	0.234	0.163

The highest average correlation was obtained for Pseudo PSNR, and Pseudo F-Measure. Pseudo F-Measure has a more stable correlation, as shown by its standard deviation. In the all datasets Pseudo NRM has the lowest average correlation.

According to the obtained results, correlation between NRM and Pseudo NRM are too low for further use and these metrics should be discarded. But all the other proposed statistically-grounded metrics can be tested and with some changes and different input parameters for algorithms.

Besides the correlation test we also evaluated how the number of classifiers influences overall correlation. Since every additional classifier has an influence on the statistically grounded metrics, we investigated what exactly happened with the correlation between the metrics when the number of classifiers is increased progressively. Fig. 2 shows how correlation between the F-Measure and Pseudo F-Measure evolves with the number of algorithms. For every quantity of classifiers, all possible combinations of mentioned binarization systems were tested. As an illustration, the data shown in Fig. 2 represents the obtained correlation evaluation for each image in the DIBCO 2011 handwritten dataset.

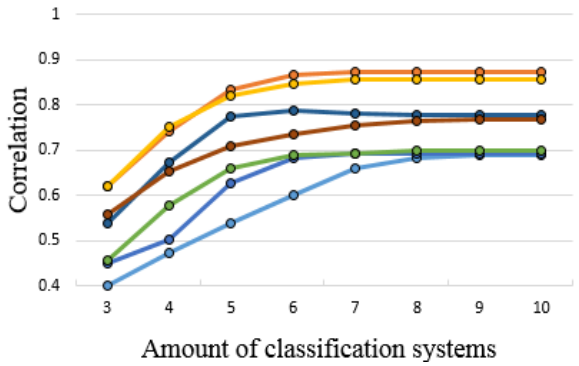

Fig. 2. Changes of average correlation with the amount of binarization algorithms (each curve represents a specific test image from DIBCO)

According to the obtained results, the more classification algorithms, the better the correlation coefficient.

VI. Conclusions And Future Work

Aiming to overcome some deficiencies of classifier performance evaluation using annotated data, this paper presents statistically grounded evaluation metrics not requiring *Ground Truth*. We have presented basic metrics for the evaluation of binary classifiers that can be computed using only statistical tools. These metrics were applied to image analysis and have been tested on DIBCO 2009 – 2013 image datasets. Their performance was compared with those of the more traditional performance metrics reported in DIBCO competitions.

The comparison between Ground Truth based and statistically grounded metrics shows high correlation for Pseudo F-Measure, Pseudo PSNR and Pseudo NCC. This means that the approach and metrics described in this paper can be used to find the best classification methods for more than half of analyzed images. In another words, we obtain better-than-random results in the selection of classifiers.

This research still needs to be completed, and there are number of questions to resolve. One of them is how to improve the statistically grounded classification metrics such that the correlation with Ground Truth-based classification metrics can be increased and whether it is possible to select the best classifier using statistical tools for a given dataset. Or, if not, how to find criteria to identify those configurations where the approach fails.

Moreover, future investigation should extend this work to non-binary classifiers and be applied to approaches in practical cases with post-classification procedures like content extraction.

References

[1] E. H. B. Smith, C. An, "Effect of 'Ground Truth' on Image Binarization", *presented at the DAS '12 Proceedings of the 10th IAPR International Workshop on Document Analysis Systems*, pp. 250-254, 2012.

[2] M.W.A. Kesiman, S. Prum, I.M.G. Sunarya, J.-C. Burie, J.-M. Ogier, "An Analysis of Ground Truth Binarized Image Variability of Palm Leaf Manuscripts", 5th Int. Conf. Image Process. Theory Tools Appl. IPTA 2015, pp. 229-233, 2015.

[3] E. H. B. Smith, "An analysis of binarization ground truthing", In Proc. Workshop on Document Analysis Systems, pp. 27-33, Boston, MA, USA, 2010.

[4] B. Lamiroy, T. Sun, "Computing precision and recall with missing or uncertain ground truth", In: Kwon, Y.-B., Ogier, J.-M. (eds.) GREC 2011. LNCS, vol. 7423, pp. 149-162. Springer, Heidelberg, 2013.

[5] B. Lamiroy, P. Pierrot, "Statistical Performance Metrics for Use with Imprecise Ground Truth" In Graphic Recognition. Current Trends and Challenges: 11th International Workshop, GREC 2015, vol. 9657 of Lecture Notes In Computer Science, Springer pp. 31-44

[6] B. Raj, R. Singh, J. Baker, "A paired test for recognizer selection with untranscribed data". ICASSP 2011: 5676-5679Acoustics, Speech and Signal Processing (ICASSP), 2011.

[7] B. Gatos, K. Ntirogiannis, and I. Pratikakis. ICDAR 2009 document image binarization contest (DIBCO 2009). In Proc. International Conference on Document Analysis and Recognition, pages 1375-1382, Barcelona, Spain, July 2009.

[8] N. Otsu, "A threshold selection method from gray-level histograms", IEEE Trans. Syst. Man Cybern., vol. SMC-9, no. 1, pp. 62-69, Jan. 1979.

[9] J. Bernsen, "Dynamic Thresholding of Grey-level Images", Proceedings of the 8th International Conference on Pattern Recognition (ICPR8), Paris, France, pp. 1251-1255, October 1986.

[10] W. Niblack, "An Introduction to Digital Image Processing", Prentice-Hall, Englewood Cliffs, NJ, pp. 115–116, 1986.

[11] Bradley, D., Roth, G.: Adaptive thresholding using the integral image. J. Graph. Tools 12(2), pp.13–21, 2007.

[12] Wolf, J-M. Jolion, "Extraction and Recognition of Artificial Text in Multimedia Documents", Pattern Analysis and Applications, 6(4):309-326, 2003.

[13] J. Kittler and J. Illingworth, "Minimum Error Thresholding," Pattern Recognition 19, 41-47, 1986.

[14] J. Sauvola, and M.Pietikainen, "Adaptive Document Image Binarization", Pattern Recognition, vol. 33, no. 2, pp. 225-236, 2000.

[15] B. Gatos, I. Pratikakis, S.J. Perantonis, Adaptive degraded document image binarization, Pattern Recognition 39 (3) pp. 317–327, 2006.

[16] B. Lamiroy, "Interpretation, Evaluation and the Semantic Gap ... What if we were on a Side-Track?" In Graphic Recognition. Current Trends and Challenges: 10th IAPR International Workshop, GREC 2013, vol. 8746, pp 213-226, Lecture Notes in Computer Science,. Springer.