IJDAR (2007) 9:59–74
DOI 10.1007/s10032-006-0033-x

ORIGINAL PAPER

A general framework for the evaluation of symbol recognition methods

E. Valveny · P. Dosch · Adam Winstanley · Yu Zhou · Su Yang · Luo Yan · Liu Wenyin · Dave Elliman · Mathieu Delalandre · Eric Trupin · Sébastien Adam · Jean-Marc Ogier

Received: 1 April 2005 / Accepted: 22 September 2006 / Published online: 18 November 2006

Abstract Performance evaluation is receiving increasing interest in graphics recognition. In this paper, we discuss some questions regarding the definition of a general framework for evaluation of symbol recognition methods. The discussion is centered on three key elements in performance evaluation: test data, evaluation metrics and protocols of evaluation. As a result of this discussion we state some general principles to be taken into account for the definition of such a framework. Finally, we describe the application of this framework to the organization of the first contest on symbol recognition in GREC'03, along with the results obtained by the participants.

Keywords Performance evaluation · Symbol recognition

1 Introduction

Performance evaluation has become an important research interest in pattern recognition during the last years. As the number of methods increases there is a need for standard protocols to compare and evaluate all these methods. The goal of evaluation should be to establish a solid knowledge of the state of the art in a given research problem, i.e., to determine the weaknesses and strengths of the proposed methods on a common and general set of input data. Performance evaluation should allow the selection of the best-suited method for a given application of the methodology under evaluation.

E. Valveny (✉)
Centre de Visió per Computador, Edifici O, Campus UAB, Bellaterra (Cerdanyola), 08193 Barcelona, Spain
e-mail: ernest@cvc.uab.es

P. Dosch
LORIA, 615, rue du jardin botanique, B.P. 101, 54602 Villers-lès-Nancy Cedex, France
e-mail: Philippe.Dosch@loria.fr

A. Winstanley · Y. Zhou
National University of Ireland, Maynooth, County Kildare, Ireland
e-mail: adam.winstanley@nuim.ie

Y. Zhou
e-mail: yuzhou@cs.nuim.ie

S. Yang
Department of Computer Science and Engineering, Fudan University, Shanghai 200433, China
e-mail: suyang@fudan.edu.cn

L. Yan · L. Wenyin
Department of Computer Science, City University of Hong Kong, Honk Kong, China
e-mail: luoyan@cs.cityu.edu.hk

L. Wenyin
e-mail: csliuwy@cityu.edu.hk

D. Elliman
University of Nottingham, Nottingham, UK
e-mail: dge@cs.nott.ac.uk

E. Trupin · S. Adam
LITIS Laboratory, Rouen University, Rouen, France
e-mail: Sebastien.Adam@univ-rouen.fr

M. Delalandre · J.-M. Ogier
L3i Laboratory, La Rochelle University, Rochelle, France
e-mail: mathieu.delalandre@univ-lr.fr

J.-M. Ogier
e-mail: jean-marc.ogier@univ-lr.fr

Following these criteria, image databases have been collected and performance metrics have been proposed for several domains and applications [6,12,18,21,29]. Several of these works deal with the evaluation of processes involved in document analysis systems, such as thinning [13], page segmentation [2], OCR [28], vectorization [22,26,27] or symbol recognition [1], among others. In fact, the general performance evaluation framework proposed in this paper is based on the work carried out for the contest on symbol recognition organized during GREC'03 [25].

Although in any domain there are always some specific constraints, we can identify three main issues that must be taken into account in the definition of any framework for performance evaluation: a common dataset, standard evaluation metrics and a protocol to handle the evaluation process. The common dataset should be as general as possible, including all kinds of variability that could be found in real data. It must contain a large number of images, each of them annotated with its corresponding ground-truth. Metrics must be objective, quantitative and accepted by the research community as a good estimate of the real performance. They must help to determine the weaknesses and strengths of each method. In many cases, it is not possible to define a single metric, but several metrics have to be defined according to different evaluation goals. The protocol must define the set of rules and formats required to run the evaluation process.

In this paper, we propose a general framework for performance evaluation of symbol recognition. For each of these issues (data, metrics and protocol), we describe the main problems and difficulties that we must face and we state the general guidelines that we have followed for the development of such a framework. Finally, we show how we have applied this framework to the organization of the GREC'03 contest on symbol recognition.

Symbol recognition is one of the main tasks in many graphics recognition systems. Symbols are key elements in all kinds of graphic documents, as they usually convey a particular meaning in the context of the application domain. Therefore, identifying and recognizing the symbols in a drawing is essential for its analysis and interpretation and a great variety of methods and approaches have been developed (see some of the surveys on symbol recognition [5,8,17] to get an overview of the current state of the art).

In fact, symbol recognition could be regarded as a particular case of shape recognition. However, there are some specific issues that should be taken into account in the definition of an evaluation framework. First, symbol recognition is not a stand-alone process. Usually, it is embedded in a whole graphics recognition system where the final goal is not only to recognize perfectly segmented images of symbols, but to *recognize and localize* the symbols in the whole document. Sometimes segmentation and recognition are completely independent processes, but sometimes they are related and performed in a single step. For evaluation, that means that we must consider two different sub-problems: recognition of segmented images of symbols and localization and recognition of symbols in a non-segmented image of a document. These two different sub-problems will be referred to as *symbol recognition* and *symbol localization*, respectively, throughout the paper. Second, sometimes, symbol recognition depends on other tasks in the graphics recognition chain (for example, binarization or vectorization). The performance of these processes can also influence the performance of symbol recognition. We should try to make the evaluation of symbol recognition independent of these other tasks. At least, the analysis of the results should be made taking into account their influence. Third, symbol recognition is applied to a wide variety of domains (architecture, electronics, engineering, flowcharts, geographic maps, music, etc.). Some methods have been designed to work only in some of these domains and have been only tested using very specific data.

Finally, if the goal of performance evaluation is to help to determine the current state-of-art of research, then, any proposal should give response to the needs of the whole research community and should be accepted by it. Therefore, in our proposal, a key point is the idea of collaborative framework. The initial proposal must be validated by the users and must be easily extended as research advances and new needs or requirements appear. Thus, our proposal relies on four desirable properties:

- public availability of data, ground-truth and metrics
- adaptability to user needs: each person must be able to select a subset of the framework to work with
- extensibility the framework must allow for new kinds of images or metrics to be easily added
- collaborative validation of data, metrics and ground-truth.

The paper is organized as follows: Sects. 2 and 3 are devoted to discuss each of the main aspects in performance evaluation, data and evaluation metrics, respectively. In Sect. 4 we describe the protocol and implementation issues of the framework. In Sect. 5 we show the application of this framework to the GREC'03 contest. Finally, in Sect. 6 we state the main conclusions and discuss the future work.

2 Data

One of the key issues in any performance evaluation scheme is the definition of a common set of test data. Running all methods on this common set will permit to obtain comparable results. This set should be generic, large, and should contain all kinds of variability of real data.

In symbol recognition, generality means including all different kinds of symbols, i.e., symbols from all applications (architecture, electronics, engineering, flowcharts, geographic maps, music, etc.) and symbols containing all types of features or primitives (lines, arcs, dashed-lines, solid regions, compound symbols, etc.). In this way, we will be able to evaluate the ability of recognition methods to work properly in any application.

On the other hand, variability can be originated by multiple sources: acquisition, degradation or manipulation of the document, handwriting, etc. All of them should be taken into account, when collecting test data in order to evaluate the robustness of recognition methods.

However, in symbol recognition many methods are specifically designed for a particular application or a particular kind of symbols under specific constraints. Therefore, it is not possible to define a single dataset containing all kinds of images. Then, following the general principle of adaptability, stated in the previous section, we propose to define several datasets, instead of a single one. Each dataset will be labeled according to the kind of images contained in it. In this way, users can select the datasets they want to use according to the properties of their method. In addition, we can generate as many datasets as required, combining all kinds of symbols and criteria of variability.

Therefore, we need to establish some criteria to classify and organize all kinds of symbols (Sect. 2.1). Then, we must also identify and categorize all kinds of variability of real images (Sect. 2.2). Finally, we will be able to discuss how to collect and generate a large amount of data and organize it according to these criteria of classification (Sect. 2.3).

2.1 Classification of symbols

In general, there are two points of view for classifying evaluation tests and their associated data [9]: technological and application. The technological point of view refers to the evaluation of methods as stand-alone processes trying to measure their response to varying methodological properties of input data and execution parameters. Datasets must be independent of the application and must differ on the kind of image features. For symbol recognition this point of view corresponds to the generic evaluation of performance independently of the application domain. Image features will be the different shape primitives that can be found in the symbols. According to the data used in the contest, we have identified three shape primitives: straight lines, arcs and solid regions. However, new primitives (for example, dashed lines, text, textured areas) could be added to the dataset if required.

On the other hand, the application point of view refers to the evaluation of methods in a particular application scenario. Different datasets will correspond to different application domains of a given method, and each dataset will only include specific data for the given application. In symbol recognition, categories refer to the different domains of application: architecture, electronics, geographic maps, engineering drawings or whatever domain we should consider.

We have used this double criteria to classify symbols in our framework. The support for it is that algorithms are usually designed using these two points of view too. Some methods are intended to be as general as possible, and work well with symbols in a wide range of applications. On the other hand, some other methods are intended to be part of a complete chain of a graphics recognition system in a particular application domain. They are specifically designed to recognize the symbols in that application.

These are the two main criteria for classifying test data. But from a more general viewpoint, we can use labels corresponding to property/value pairs. The property can refer to the application domain, primitives, origin, etc., while values are occurrences of these properties (respectively, architecture/electronic/. . ., segments/arcs and segments/. . ., CAD design/sketch/. . .). This provides a general labeling system which can be easily extended, allowing to define as much data as needed.

Therefore, we will assign at least two categories of labels to each symbol: one with the domain of the symbol and the other with the set of primitives composing it. Each dataset is also labeled in the same way according to the symbols included in it. With this organization each user can select those datasets that fit the features of the method under evaluation. In addition, new categories of data can be easily added or modified and therefore, the framework can evolve according to research needs. In Fig. 1 we can see several examples of images classified according to both points of view. Note that each symbol can be included in several categories.

2.2 Variability of symbol images

Robustness to image degradation is essential for the development of generic algorithms. Then, a framework

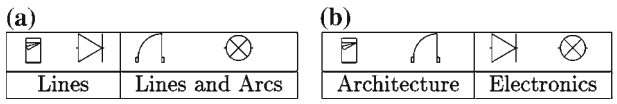

Fig. 1 Classification of the same images according to the two points of view: **a** technological, **b** application

for performance evaluation must include all kinds of degradation in the test data. Besides, images should be ranked according to the degree of degradation in order to be able to determine whether the performance decreases as the difficulty of images increases.

In general, we can distinguish four sources of variability in symbol recognition:

- acquisition parameters: acquisition device (scanner, camera or online device) and acquisition resolution
- global transformations: global skew of the document, rotation and scaling of symbols
- binary noise: degradation of old documents, photocopies, faxes and binarization errors.
- Shape transformations: missing or extra primitives (due to segmentation errors) and shape deformations due to hand-drawing.

We need to guarantee that all these types of degradations are included in the common dataset. We will generate different datasets corresponding to each kind and degree of transformation and to selected combinations of them. Each dataset will be labeled accordingly too.

2.3 Generation of test data

According to the principles stated in previous sections we need to collect a large number of images. These images will be organized into several datasets, including all kinds of symbols described in Sect. 2.1 and all types of variability identified in Sect. 2.2. In addition, images must be labeled with the ground-truth, i.e., the expected result. We have to collect segmented images of isolated symbols, but also non-segmented images of documents in order to evaluate both symbol recognition and symbol localization, as stated in Sect. 1.

There are basically two possibilities for collecting test data: to use real data or to generate synthetic data. In the following of this section, first, we will discuss the advantages and drawbacks of each approach and how we use them in our framework. Then, we will consider some other specific issues related to the generation of data for evaluation of symbol recognition.

2.3.1 Real data

Clearly, the main advantage of using real data is that it permits to evaluate the algorithms with the same kind of images as for real applications. Then, evaluation will be a very good estimate of performance in real situations. However, manually collecting a large number of real images is a great effort, unaffordable in many cases. The task of annotating images with their corresponding ground-truth is also time-consuming, and errors can easily be introduced. Another disadvantage is the difficulty of collecting images with all kinds of transformations and noise. Besides, it is not easy to quantify the degree of noise in a real image. Then, it is not possible to define a ranking of difficulty of images according to the degree of noise.

2.3.2 Synthetic data

As an alternative, we can develop automatic methods to generate synthetic data. Clearly, the main advantage is that it allows to generate as many images as necessary, and the annotation of images with the ground-truth is also automatic. Then, manual effort is reduced. However, we need to devote research effort to the development of models and methods able to generate images resembling real ones with all possibilities of noise and transformations. This is not an straightforward task in many cases although several works have been done in related fields of document analysis [3,11,15,16]. Images generated using these methods will be easily classified according to the type and degree of noise or degradation applied, permitting to assess the reduction in performance with increasing degrees of image degradation.

We argue that both types of images are useful in a general framework for performance evaluation of symbol recognition. We believe that real images are the best test for assessing performance in symbol localization. It is really difficult to develop automatic methods to generate non-segmented images of complete graphic documents. Besides, as we can find many symbols in a single graphic document, not many images are required. The problem can be the annotation of images with the ground-truth. We discuss it in Sect. 3.3.

On the other hand, synthetic images are the only way to perform evaluation tests with large sets of segmented images taking into account all degrees of degradation and variation. In this case, many images are required and it is easier to develop methods for their generation. In our framework we have developed methods for the generation of global transformations, binary noise (based on Kanungo's method [15] and shape transformation (based on active shape models [25]).

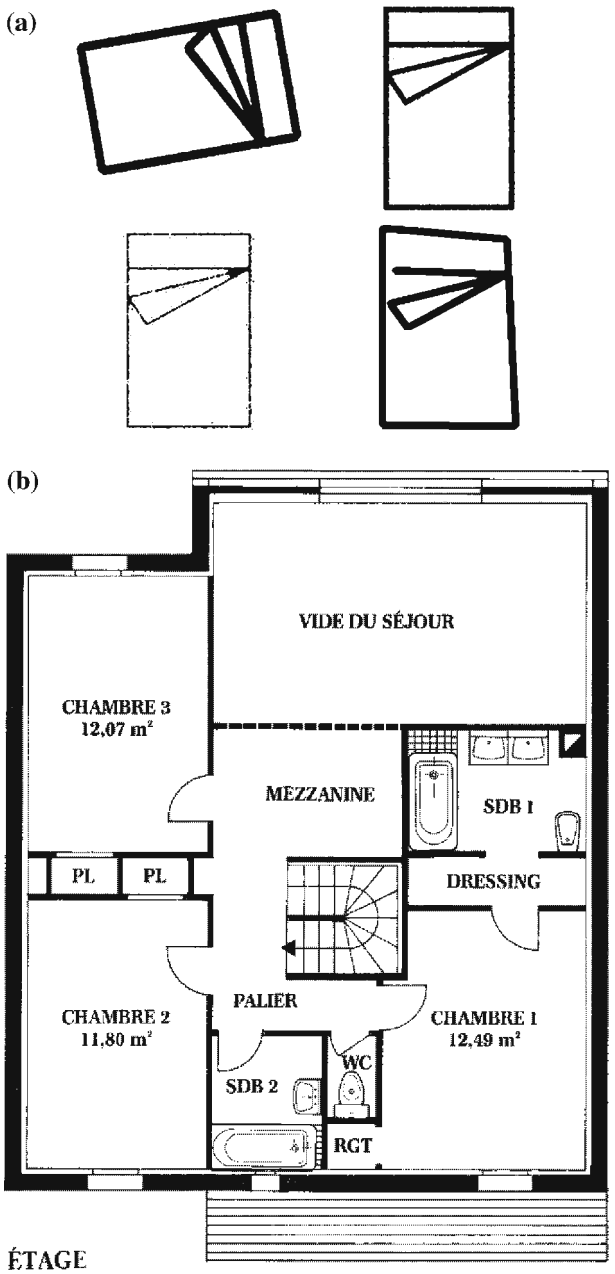

Fig. 2 Generation of data: **a** synthetic images, **b** real images

Figure 2 shows both synthetic and real images for symbol recognition.

2.3.3 Specific issues

In addition, we have to take into account two other specific issues of symbol recognition when generating test data.

- *Relation to vectorization*: As explained in Sect. 1 symbol recognition is simply one task in the graphics recognition chain. Vectorization is usually performed as a previous step for recognition and then, many symbol recognition methods work directly on the vectorial representation of the image. The problem is that, although there is not an optimal vectorization method, the result of vectorization can influence the performance of recognition. Then, apart from a raster representation of images, we must also provide images in a common vectorial format so that all methods can use the same vectorial data and recognition results are not influenced by the selected vectorization method. For images that can be automatically generated in vectorial format, we can provide images in their ideal vectorial representation, without need for applying any vectorization method. If not possible (for example, for real images of for synthetic images with binary degradations), we should apply different standard vectorization methods to the raster image.
- *The problem of scalability*: One of the problems in symbol recognition [17] concerns scalability: many methods work well with a limited number of symbol models, but their performance decrease when the number of symbols is very large (hundreds or thousands of symbols). One of the goals of the evaluation of symbol recognition must be to assess the robustness of methods with a large number of symbols. Then, for each kind of test several datasets with an increasing number of symbols will be generated.

3 Performance evaluation

3.1 Objectives

In some pattern recognition fields, the main goal of evaluation is the definition of a global measure that permits to determine the "best" method on a standard and common dataset. However, it seems difficult to follow the same approach for symbol recognition. As we have stated in previous sections, performance of symbol recognition depends on many factors and it is not realistic trying to define a single measure and dataset taking into account all of them. Then, as symbol recognition remains an active research domain, it seems more interesting to focus on analyzing and understanding the strengths and the weaknesses of the existing methods. This will be the main goal of the proposed evaluation framework.

In this context, evaluation relies on three issues: first, the definition of a number of standard datasets, covering the full range of variability, as discussed in Sect. 2. Second, the definition of a set of measures, each of them aiming at evaluating a specific aspect of performance.

This will be discussed in Sect. 3.2. The definition of metrics is highly related to the definition of the ground-truth. This point will be developed in Sect. 3.3. Third, the analysis of the results after calculating all the measures over all the datasets, in order to draw conclusions on the strengths and weaknesses of each method (Sect. 3.4).

3.2 Metrics

In the last years, several graphics recognition contests have been organized, notably in the framework of the International Workshop on Graphics Recognition (GREC). As a result of this effort, several metrics and protocols have been developed [14,22,26], with more or less success, as sometimes, they favor the properties of some of the contestant methods.

A similar work has to be done for symbol recognition: what is the measure that permits to say that a given symbol recognition method is good? Clearly, the answer will be different for each of the two sub-problems identified in Sect. 1: *symbol recognition* and *symbol localization*. In the first case, for the recognition of isolated symbols, it can be enough to count the number of correctly recognized symbols. But, in the second case, other information, such as location, orientation and scale of symbols should also be considered. Thus, in the following, we will discuss different metrics for each of these sub-problems.

3.2.1 Symbol recognition

It seems clear that the basic metric for symbol recognition should be to test if the recognized symbol matches the test symbol according to the ground-truth. Thus, the recognition rate is the main evaluation criteria. This was the simple approach used in the GREC'03 contest. Because of the wide number of open questions regarding performance evaluation of symbol recognition, we decided, in a first time, to consider only the basic features in order to advance in a better understanding of all issues involved in it.

However, we believe that this criteria could be complemented with other measures, in order to get a deeper analysis of recognition methods, taking into account other evaluation aspects. For example,

- The recognition rate, considering second or third candidates, if this information is provided by some methods.
- The orientation and scale of the symbol: we could complete the recognition rate with a measure of the accuracy in recovering the orientation and scale of the symbol. This measure can be based on the difference between the orientation and scale provided by the recognition method and the ground-truth.
- The computation time: we propose to use the average time per image. This metric will allow to compare the results on tests with different number of images or symbols. However, to be comparable, all recognition methods should be run on the same machine under the same conditions. That should be considered in the definition of the protocol (Sect. 4.2).
- Scalability, i.e., how the performance degrades as the number of symbol models increases. We can measure it according to the degradation of recognition rates or according to the computation time.

3.2.2 Symbol localization

In the best of our knowledge, no performance evaluation has ever been organized on symbol localization. For this task, the problem of defining accurate metrics is harder than in the case of symbol recognition. We have to face two issues: the representation of the symbols, and the definition of the metric itself.

The representation of a symbol (in the ground-truth as well as in the recognition result) must include not only an identifying label (as in the case of symbol recognition), but also the location of the symbol. The problem is that it is not easy to define a single representation of the location of a symbol. The best representation will depend on the kind of method. For example, if a recognition method works on the raster representation of a symbol, the symbol location has to be computed with respect to the related set of pixels. But if a recognition method works on the vectorial representation of the symbol, its location has to be computed with respect to the involved set of vectorial primitives, maybe taking into account some attributes of these primitives, such as thickness. Clearly, both representations do not have to be equal.

In fact, we argue that the representation of the location of a symbol must be unique and independent of the kind of method or image format, as the definition of multiple representations arise the following issues:

- Multiple metrics have to be defined as the definition of the metric depends on the representation of the symbols. This can permit to define more accurate metrics but also requires to take into account all possibilities.
- Multiple representations also lead to the definition of multiple ground-truth for the same data.

- Multiple metrics and multiple ground-truth then lead to multiple performance analysis as it will be difficult to compare results evaluated with different metrics.

As a first approach for representing the location of a symbol, we propose the use of basic including rectangles, that enclose symbols, as described by Mariano et al. [20]. This representation seems to be simple and efficient. These rectangles can even be defined as bounding-boxes.

Then, the metric between a ground-truth symbol and a result symbol can be based on the percentage of overlapping between their including rectangles, in the case that their associated labels match. Otherwise, the similarity value will be 0. This metric permits to work at the desired level of accuracy. We can fix a threshold so that only symbols with a percentage of overlapping above this threshold are considered as recognized. In this way, defining several thresholds, we can obtain different recognition results at different levels of accuracy.

In order to combine the results of the metric obtained for every symbol in the image, we propose to adopt a metric similar to the one used during the ICDAR'03 conference on the robust reading competition [19] for the text recognition in everyday scenes. The definition principles are based on the fact that the metric must favor the most pertinent applications, and penalize trivial solutions, like the definition of a single bounding-box which fully overlaps the image, or the definition of an excessive large number of bounding-boxes.

So the proposed metric is based on the notions of *precision* and *recall*. For a given test, let T be the number of targets belonging to the ground-truth, and R the set of results supplied by an application. The number of exact results is called e. The precision p is then defined as the number of exact results divided by the number of results:

$$p = \frac{e}{|R|}.$$

Thus, the applications that overestimate the number of results are penalized by a little precision score. The recall r is defined as the number of exact results divided by the number of targets:

$$r = \frac{e}{|T|}.$$

Thus, the applications that underestimate the number of results are penalized by a little recall score. The precision and the recall may then be combined, if needed, to determine the global score s, expressing the recognition rate:

$$s = \frac{2}{(1/p) + (1/r)}.$$

3.3 Ground-truth

As said above, the definition of the ground-truth depends basically on the representation of the symbols. Once again, we have to distinguish between the definition of the ground-truth for symbol recognition and for symbol localization.

If we consider symbol recognition, where only segmented symbols are involved, ground-truthing can be a simple task. It basically consists of determining the label of the symbol and this can be easily done by an human operator and even, more easily by an automatic method of image generation. If we also want to take into account the accuracy in orientation and scale, we must include this information in the labeling of the symbol too. But this can be easily done with an automatic method of image generation.

However, if we consider symbol localization, ground-truthing is more difficult. In this case, both the label and the location of the symbol have to be defined. According to the single proposed metric (see Sect. 3.2), the definition of the ground-truth is also unique, and then easier and more realistic to manage.

Although the representation of the symbol gives a theoretical and concrete framework for the definition of the ground-truth, some differences can exist between the theoretical definition and the real definition of a given ground-truth. Indeed, the bounding-box defined by one person for a given symbol could appear misplaced to another person. Thus, there is a part of personal and subjective interpretation in the definition of the ground-truth.

This point can be a serious problem, as the ground-truth has to be accepted by the whole community to be fully considered as a reference. To address this issue, we are fully convinced that a collaborative framework is required, as already pointed out in Sect. 1.

The basic idea is to involve a ground-truth designer and some ground-truth validators for a given ground-truth. Meanwhile, a ground-truth definition can be modified if it is not satisfactory. Of course, a ground-truth designer of some test data cannot be the ground-truth validator of the same test data too. Once a ground-truth is validated by some people, say two or three, then, it can be considered valid. This organization could be compared to a review process for a scientific conference. Obviously, this organization is easier to implement if a collaborative tool is available, as the associated workflow is crucial. This tool includes the following features:

- General ground-truthing functionalities: images visualization (raster, vectorial), bounding-box definition, label definition ...
- Directly interfaces with the database implementing the information system containing all information required for performance evaluation:
  - information about the data: models of symbols, test data and related ground-truthing.
  - information about users involved in the evaluation: their role and corresponding access privileges (ground-truth design and validation, data contributor ...)
- The collaborative tool must be unique, in order to be used in good conditions by all people involved in the ground-truthing process. This implies that it has to be available for a sufficient number of platforms and ensures that all people work with the same environment or references.

We want to point out that these principles and this framework are a priori necessary in order to ensure that test data, as well as their associated ground-truth, are considered as valid by the whole community, and not by only one person. All the performance evaluation process relies on this assertion.

3.4 Analysis of the results

The results of the participants have to be analyzed in order to determine the objectives of such a performance evaluation campaign: the understanding of the strengths and the weaknesses of the existing methods. This analysis must be done with respect to the considered categories of data, the number of model symbols involved and several other interesting criteria.

Independently of this large number of criteria, we would point out that basically the analysis can be led from the data point of view (data based), as well as from the methods point of view (methods based). Indeed, if it is interesting to understand what are the methods giving good results with a lot of data, it is also interesting to understand what are the data difficult to recognize with respect to the several recognition approaches. The interest of a performance evaluation campaign is guided by these two points of view.

Based on the metric that has been defined for symbol recognition, we propose to define an index that permits to perform the analysis of the results from different points of view. This index is a measure of the degradation of the performance along a set of tests with an increasing level of difficulty. Let r_0 be the recognition rate for the test acting as the reference test (it should be the "easiest" test in the series). Then the degradation of performance for a given test i is defined as

$$d_i = \frac{r_0 - r_i}{r_0}.$$

This index gives the measure of how the original performance degrades when some kind of degradation is applied to the original images. As the index is normalized by the original recognition rate it provides a good estimate of the loss of performance as it does not depend on the recognition rate for ideal images.

In this way, we can measure the robustness of recognition methods to several properties, such as scalability or degradation. We simply need to define a series of tests with an increasing number of symbols (for scalability) or with different levels of degradation and compute the degradation index for every test. Some examples of the application of this index to the analysis of the results will be shown in Sect. 5.

4 Implementation

4.1 Introduction

The implementation of any performance evaluation system requires the definition of a set of tools and protocols in order to execute the tests, exchange information between the participants and the organizers and manage all the information about test data and results. This set of tools and protocols must rely on the general concepts stated in Sect. 1, such as the public availability of data, the adaptation to user requirements and the simplicity of management.

Among all these issues, in the remainder of this section we will discuss the main ideas regarding protocols and formats (Sect. 4.2), the organization of datasets (Sect. 4.3) and the general architecture of the system (Sect. 4.4).

4.2 Protocols and formats

Whatever the evaluation criteria and data, an evaluation framework must provide formats and tools allowing to exchange information about models, tests and results [24]. In performance evaluation of symbol recognition, the first issue is about the format of images. One basic assumption to be made is that the format of images must not degrade the original image and must be freely available for all participants. As there are methods working on raster binary images and methods working on vectorial images, whenever it is possible, we have to

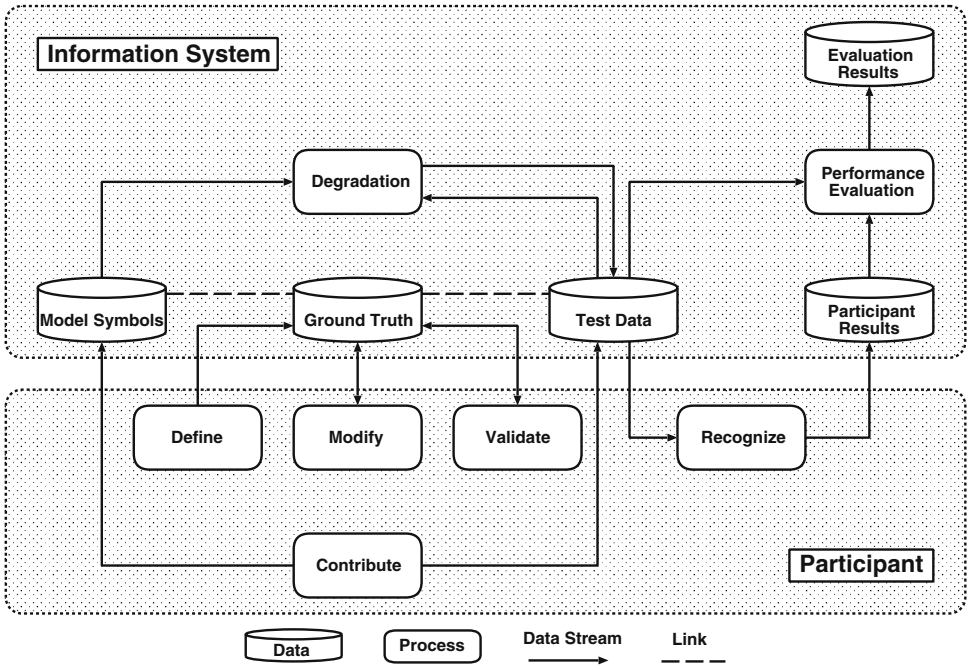

Fig. 3 Overview of the discribed performance evalution system

provide test images in both formats. Raster images are not a big problem as there are a lot of very popular solutions (such as TIFF, BMP and PNG). On the vectorial side, some "standard" formats exist, such as DXF or more recently SVG, but they are complex to manage. Thus, we have decided to use a simpler vectorial representation, the VEC format proposed by Chhabra and Phillips [4]. This simple format have already been used in other contests on graphics recognition (vectorization and arc detection) and therefore, it is already known by the symbol recognition community. Moreover, the simplicity of its definition would permit to eventually extend it, if required.

To manage the contest, several other file formats are required to precisely describe the tests, the results and the ground-truth. In this case, the choice of the format is a question of finding the best compromise that permits to express all the information that is required without obliging the participant methods to interface with too complex formats. We have found that XML fulfills these requirements as it is a flexible and standard format, allowing to easily describe complex information. Moreover, the use of a DTD or a scheme can help to normalize the data, avoiding description problems or confusions, and associated with the XSLT style-sheets, it allows the extraction and filtering of data that can be automatically processed, both for participants and organizers. Examples of these XML files can be seen in Figs. 3 and 4.

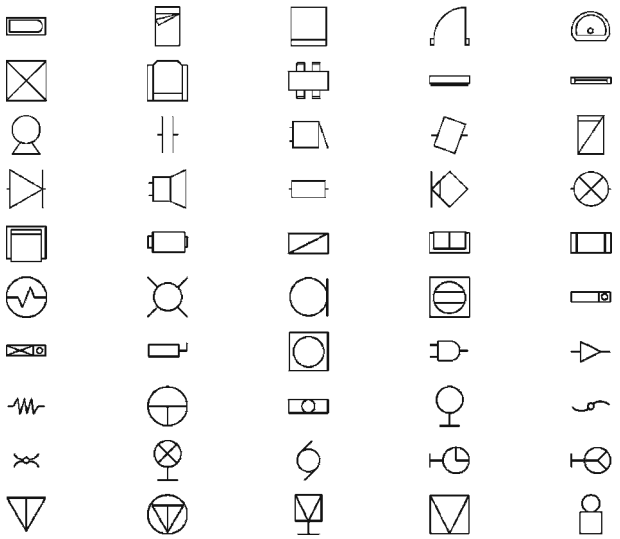

Fig. 4 Fifty symbols used in Contest

Another important issue is the protocol for execution of the tests. Following the principle of adaptability to user requirements, the basic idea must be to give each participant the possibility to choose which tests he want to compete in, according to the features of his method. To achieve this point, each test has to be considered as a stand-alone part and described with an independent XML file as explained in the next section. This principle is useful in some other situations. Thus, if a program crashes during a test, it is able to run the other tests.

The model that we have selected for the execution of the tests is a distributed model: each participant can take a file describing a test, execute it locally and then, provide the XML file with the results to the organizers. This option gives the maximum freedom to the users, for example regarding the platform of development or the interface of the recognition method. This is coherent with the general principles of the framework, but it can also have some drawbacks as the organizers do not have complete control on the development of evaluation and on some of the results. For example results regarding computation time are not fully comparable.

Finally, we want to point out that the availability of the framework (formats, data, etc.) is very important. In the context of performance evaluation, information about formats and data is required to prepare the methods for running the tests and for learning purposes.

4.3 Organization of datasets

A general framework for performance evaluation must include a very large number of datasets, taking into account all the variability described in previous sections. In order to manage this volume of datasets, we have to organize and classify them according to their properties. We will achieve this goal in a double way. On one hand, internally, we will store all information of every test in the information system that supports the evaluation framework and is described in the next section. On the other hand, externally, we will make it public to the participants by providing an XML description file for every test, as can be seen in Fig. 3. This file contains all the information that a participant has to know about a test:

- the name of images
- the ground-truth for each image (for training sets only)
- the category of symbols (as described in Sect. 2.1) from technological and application point of view
- the number of symbols involved in the dataset (for scalability issues)
- supported formats for images in the test
- whether the test corresponds to segmented or non-segmented images
- whether the test includes real or synthetic images
- whether the image acquisition is online or offline
- the type and degree of degradation applied to the data.

This organization allows to describe each test, so its associated properties are known. In this way, each participant can select the tests with the properties that fit to the method being evaluated. Moreover, it facilitates the analysis of the results, as it allows to organize the analysis according to the properties of the tests.

4.4 Information system

In order to manage all this framework, we propose to implement an information system supporting all required features. This information system must be implemented on the organizer's side, but it must be of public access and available through the Web with standard navigation tools. It plays the role of a public repository where any user (participant, organizer, ground-truth validator) can find all the required information about the evaluation process. However, the users are not tied to the implementation of the information system as the access is done through the web and all the exchange of information through the XML files that have been described in Sect. 4.2. Providing public access to all the information about data stored in the information system permits to set up a continuous evaluation framework. Evaluation does not depend on some predefined milestones, such as the organization of specific contests, but any user can, at any moment, download a set of tests, run a given method on them and provide the results back to the organizers. In this way we obtain the maximum flexibility for evaluation of current research.

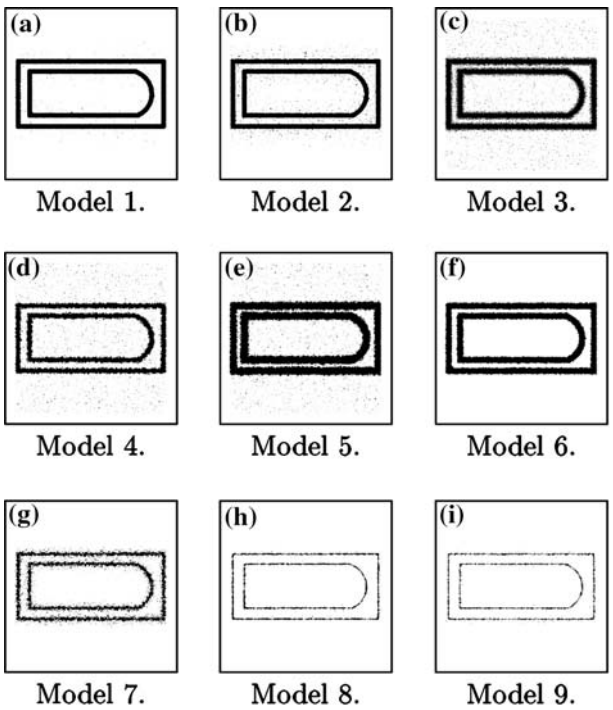

Fig. 5 Samples of some degraded images generated using the kanungo method for each model of degradation used

An overview of the system is presented in Fig. 5. Of course, the processes associated to the "participants" are related to all kinds of participants (contributors, ground-truth designers, contest participants ...) and some constraints are associated to the system. In particular, a participant cannot validate a ground-truth he has defined before, he cannot get his own test data (at least if it has not been degraded before), etc. Our aim is to point out that collaborative aspects must be taken into account from the beginning of the design of such a system.

5 Application of the framework: contest on symbol recognition at GREC'03

In this section we will show an example of application of the general framework presented before used in the *First Contest on Symbol Recognition* held during GREC'03. In this section we will explain how we have defined the two main issues involved in evaluation systems: data and metrics. We will also show the results obtained by the participants in the contest.

5.1 Data

The first decision concerned which symbols we were going to use in the contest and how to classify and organize them. For this first edition of the contest, we selected 50 symbols from two domains: architecture and electronics. All symbols were composed of at most two graphical primitives: lines and arcs. Then, according to the classification introduced in Sect. 2.1 we have used two features at the technological level (lines and arcs) and two categories at the application level (architecture and electronics) which have been used to classify test data. In Fig. 6 we can see all the symbols used in the contest.

We decided to use only synthetic data since it was easier to have a lot of well-organized images. Regarding the variability of data we worked with five categories of images: ideal data, images with aspect transformation (rotation and scaling), images with binary noise, images with shape distortions and images combining binary noise and shape distortion. We used the degradation model of Kanungo et al. [15] to generate nine different models of binary noise, and we defined a shape-distortion model based on *Active Shape Models* [7] to simulate hand-drawn images. Figures 7 and 8 show some examples of images with binary noise and shape degradation, respectively.

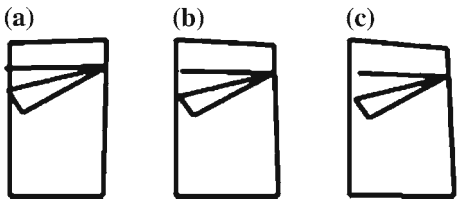

Fig. 6 Examples of increasing levels of vectorial distontion

Concerning specific issues of symbol recognition, we only used segmented images, so that only recognition was evaluated and not the ability to segment. Whenever possible, we provided both binary and vectorial versions of images. We used ideal vectorial representation when it could be automatically generated by the generation model. Therefore, for images with binary noise, only the binary representation was available as we did not apply any vectorization method to noisy binary images. Finally, we defined three different sets of symbols, with 5, 20 and 50 symbols each, to test the robustness of methods to scalability.

With all these combinations we generated a total number of 72 different tests of data. For each test, we provided a description file to the participants with the specification of symbols and images included in the test. Besides, we generated an XML file (Fig. 3) for each test, describing all the properties of the test, along with the ground-truth. Finally, participants generated an XML file (Fig. 4) with the description of the results obtained by their method for each test. Both kinds of XML files were imported to the contest database allowing for automatic comparison of the results with the ground-truth and automatic generation of recognition rates for each method and test.

5.2 Metrics

In this case, the definition of the metrics was very simple. We only worked with non-segmented images and, therefore, the only result of the application of a symbol recognition method was the label of the symbol identified in the image. Then, the metric simply consists of a recognition rate for each method and test, without taking into account the rejection.

5.3 Results

Five methods took part in the contest, although not all of them could run all the tests, due to the properties of their methods. The five participants were groups from the following institutions: University of Rouen—La Rochelle, National University of Ireland—Maynooth, City University of Hong Kong, University of Nottingham and Fudan Universty.

Fig. 7 Examples of XML file for test description

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE test SYSTEM "GRECTestSpecifications.dtd">
<?xml-stylesheet type="text/xsl" href="GRECOrgan2Particp.xsl"?>
<test format="bitmap"
      segmentation="true"
      applicationdomain="architecture">
  <testname>degrad-level1-m1</testname>
  <modelspath>models</modelspath>
  <imagespath>degrad-level1-m1</imagespath>
  <noise>
    <degradation type="kanungo">
      <noiseparam name="a0" value="0.5"/>
      <noiseparam name="a" value="0.5"/>
      <noiseparam name="b0" value="0.005"/>
      <noiseparam name="b" value="0.001"/>
      <noiseparam name="eta" value="0.0"/>
      <noiseparam name="ero" value="1.0"/>
    </degradation>
  </noise>
  <model name="ArchitecturalB.vec" id="m1"/>
  <model name="ArchitecturalC.vec" id="m2"/>
  <model name="ArchitecturalF.vec" id="m3"/>
  <model name="ArchitecturalG.vec" id="m4"/>
  <model name="ArchitecturalJ.vec" id="m5"/>
  <testimage name="image1.tif">
    <refmodel ref="m1"/>
  </testimage>
  <testimage name="image2.tif">
    <refmodel ref="m5"/>
  </testimage>
  <testimage name="image3.tif">
    <refmodel ref="m1"/>
  </testimage>
  <testimage name="image4.tif">
    <refmodel ref="m3"/>
  </testimage>
  ...
</test>
```

In Figs. 9, 10, 11, 12, 13, 14, 15, 16, we can see the results obtained by each of the methods in the tests they took part in. Figure 9 shows the results with ideal images of the symbols for the sets of 5, 20 and 50 symbols. It shows how the methods are able to discriminate among a large number of symbols. In Fig. 10 we can find the results for rotated and scaled images (for the set of 5, 20 and 50 symbols too).

Figure 11 contains the results with binary degraded images. In this case, only two methods were run on all the images and, therefore, only the results for these two methods are included. For each of the nine models of degradation the results with 5, 20 and 50 symbols are shown. In order to provide a more detailed analysis of the results with degradation we have also generated Fig. 12. In this figure we apply the degradation index defined in Sect. 3.4 to the nine models of binary degradation with the set of 50 symbols. The reference recognition rate for computing the index is the recognition rate for ideal images. This index clearly shows that for all models of degradation the method by the Fudan University is more robust to degradation than the method by the City University of Hong Kong.

Figures 13 and 14 show the results for images with vectorial distortion (for three levels of distortion) and with a combination of vectorial distortion and binary degradation.

In order to evaluate more precisely the scalability of methods we have included Fig. 15. This figure has been generated taking, for each method, the mean of recognition rates for all tests with 5 symbols, for all tests with 20 symbols and for all tests with 50 symbols. In this way, we can get a measure of the global scalability of each method. In Fig. 15a we can see the absolute recognition rates, while in Fig. 15b we have the degradation index defined in Sect. 3.4 applied to scalability. It is clear that this index helps to see the robustness of each method as the number of symbol increases.

Fig. 8 Examples of XML file for discription of results

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE testresult SYSTEM "GRECParticipantResults.dtd">
<testresult>
  <testname>testgrec</testname>
  <participant>participant1</participant>
  <imageresult>
    <imagename>
      testgrec-image1.tif
    </imagename>
    <symbol>
      <symbolname>ArchitecturalB.vec</symbolname>
      <location x1="35" y1="65" x2="67" y2="108"/>
      <orientation>34</orientation>
      <confidencerate>0.9885</confidencerate>
    </symbol>
  </imageresult>
  <imageresult>
    <imagename>
      testgrec-image2.tif
    </imagename>
    <symbol>
      <symbolname>ElectricalC.vec</symbolname>
      <location x1="108" y1="2003" x2="54" y2="21"/>
    </symbol>
  </imageresult>
  <imageresult>
    <imagename>
      testgrec-image4.tif
    </imagename>
    <symbol>
      <symbolname>ArchitecturalF.vec</symbolname>
      <location x1="30" y1="77" x2="165" y2="245"/>
    </symbol>
    <symbol>
      <symbolname>ArchitecturalF.vec</symbolname>
      <location x1="450" y1="479" x2="35" y2="88"/>
    </symbol>
  </imageresult>
</testresult>
```

Finally, in Fig. 16 we can see the computation time for every kind of test for sets with 5, 20 and 50 symbols. Only the method by the City University of Hong Kong reported results about the computation time. As expected, computation time increases as the number of symbols in the dataset increases too.

From these results we can draw some general conclusions:

- As expected, performance decreases when the number of symbols increase, even with ideal images.
- In general, methods can handle well the images with rotation or scaling. However, the performance degrades when both transformations are combined.
- There are no significant differences in the performance for the nine models of binary degradation.
- Methods are robust to the kind of shape deformations generated by the model of deformation.

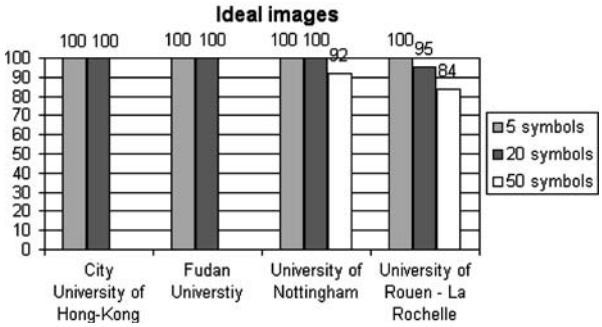

Fig. 9 Recognition rates (in the *y*-axis) of each participant method (in *x*-axis) for ideal tests

A more detailed discussion of these results can be found in the report on the GREC'03 contest [25].

Later, some of the groups have done further work on their methods and have obtained and published improved results [10].

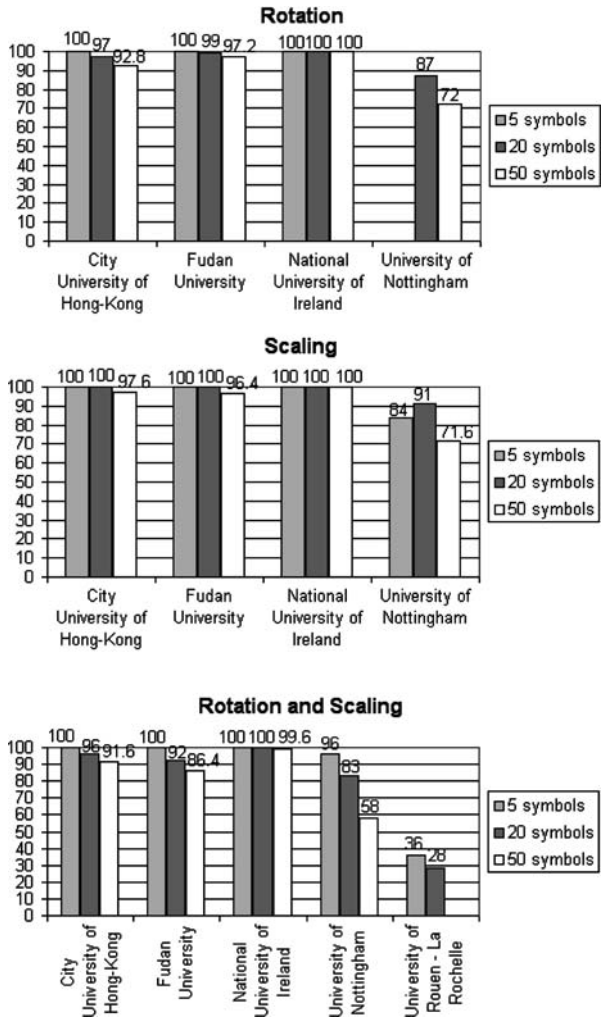

Fig. 10 Recognition rates (in the *y*-axis) of each participant method (in *x*-axis) for tests with rotation, scaling and combination of rotation and scaling

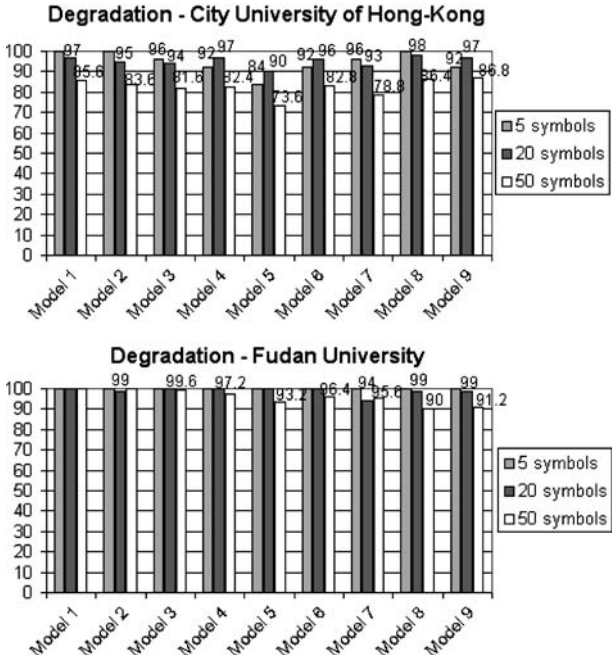

Fig. 11 Recognition rates (in the *y*-axis) for tests with the nine models of degradation (in *x*-axis) for methods by the City University of Hong Kong and the Fudan University

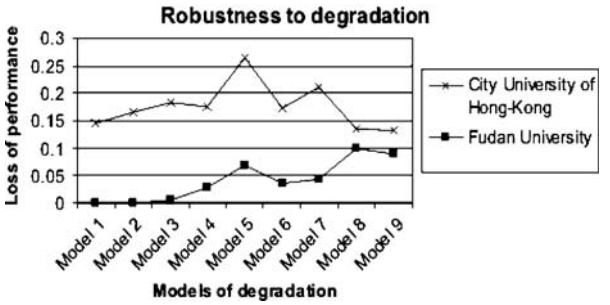

Fig. 12 Measure of robustness to degradation for the nine models of degradation with 50 symbols

6 Conclusion and future work

We have presented a general framework for performance evaluation of symbol recognition methods. This framework relies on some general principles that could also be applied to other similar performance evaluation tasks in the domain of graphics recognition and pattern recognition. These general principles arise from the discussion about the two main issues concerning any performance evaluation task: data and evaluation.

Concerning data, the framework relies on the classification of input data according to two different points of view: methodological—based on image features and application—based on the application scenario. This classification permits to define many different datasets for all possible kinds of input data. Regarding data generation we have stated the importance of using both real and synthetic images, including all types of noise and distortion. We have introduced a possible classification of distortion types and remarked the importance of including in the framework models and methods for automatic generation of degraded images.

Concerning evaluation, we have defined several metrics for symbol recognition and symbol location. Each metric gives response to different goals of performance evaluation.

In addition, one of the key ideas in the proposed framework is that of collaborative work so that the framework can be validated by the research community, and evolve according to its needs. Following this idea, a public and collaborative environment for performance evaluation of symbol recognition methods, ÉPEIRES,[1]

[1] http://www.epeires.org

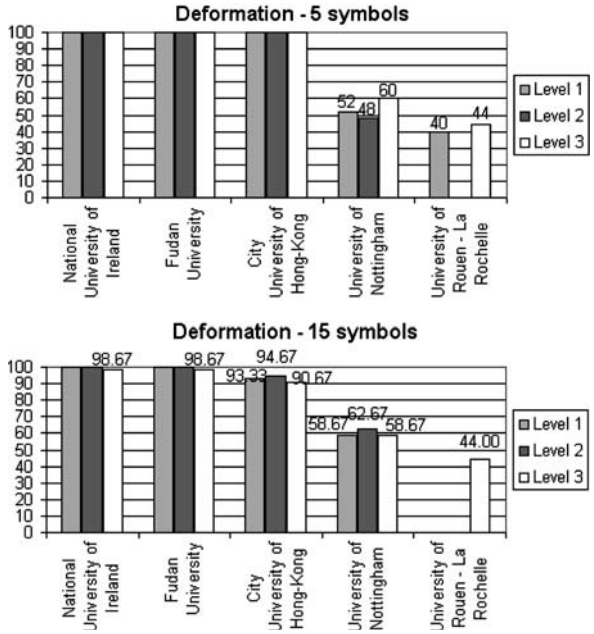

Fig. 13 Recognition rates (in the *y*-axis) of each participant method (in *x*-axis) for tests with deformation for both sets of symbols

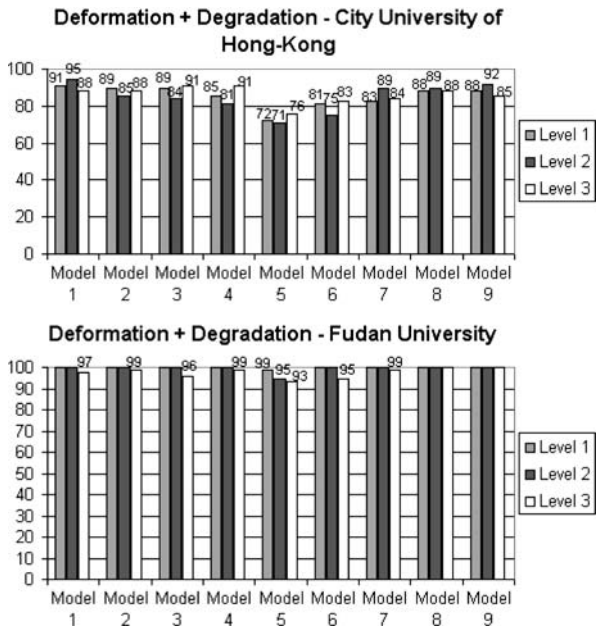

Fig. 14 Recognition rates (in the *y*-axis) for tests with the nine models of degradation (in *x*-axis) and three levels of degradation for methods by the City University of Hong Kong and the Fudan University

is currently under development. We hope that this environment will supply all data and resources needed by the symbol recognition community for evaluation purposes. All interested people are urged to use and to contribute to this environment.

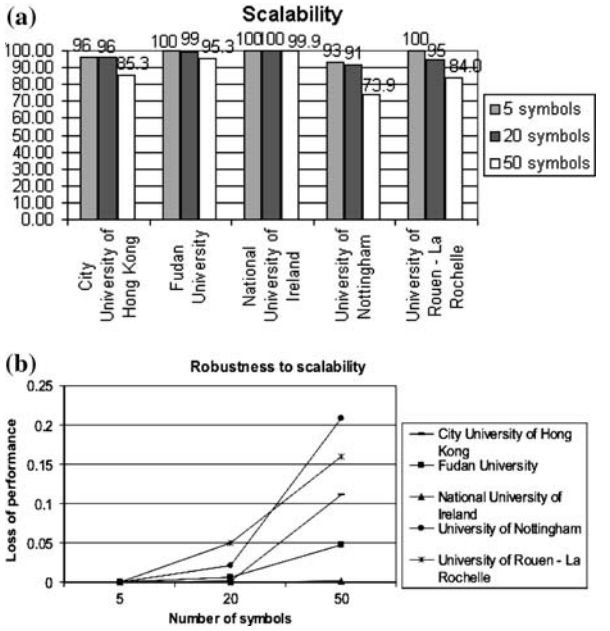

Fig. 15 **a** Evolution of recognition rates (in the *y*-axis) of each participant method (in *x*-axis) for tests with increasing number of symbols (5,20 and 50). **b** Measure of robustness to scalability for each participant method

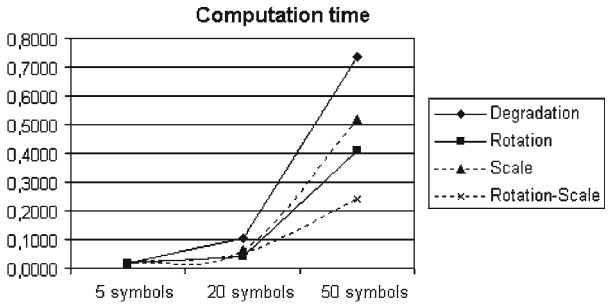

Fig. 16 Evolution of the computation time with the method by the City University of Hong Kong with an increasing number of symbol for each kind of test

Finally, we have described how these general principles have been used in the first international contest on symbol recognition, held during GREC'03. Currently, we are working on the extension of the framework for the next editions of the contest. In it, we plan to add real images with non-segmented symbols and, therefore, we will need to include the new metrics for symbol localization, as discussed in this paper.

Acknowledgments The contest organizers would like to acknowledge all participants of the first contest of performance evaluation of symbol recognition methods, as well as the organizers of the GREC workshop for the promotion and the opportunity given in these contests. The work of Luo Yan and Liu Wenyin was fully supported by grants from the City University of Hong Kong (Project No. 7001771 and 7001842) The work of E. Valveny was partially supported by CICYT TIC2003-09291, Spain.

References

1. Aksoy, S., Ye, M., Schauf, M., Song, M., Wang, Y., Haralick, R., Parker, J., Pivovarov, J., Royko, D., Sun, C., Farneboock, G.: Algorithm performance contest. In: Proceedings of 15th International Conference on Pattern Recognition, vol. 4, pp. 870–876, Barcelona, Spain (2000)
2. Antonacopoulos, A., Gatos, B., Karatzas, D.: ICDAR 2003 page segmentation competition. In: Proceedings of 7th International Conference on Document Analysis and Recognition, Edinburgh (Scotland, UK), pp. 688–689 (2003)
3. Baird, H.S.: The state of the art of document image degradation modeling. In: Proceedings of 4th IAPR International Workshop on Document Analysis Systems, Rio de Janeiro (Brazil) (2000)
4. Chhabra, A., Phillips, I.T.: The 2nd international graphics recognition contest—raster to vector conversion: a report. In: Tombre, K., Chhabra, A.K. (eds.): Graphics Recognition—Algorithms and Systems. Lecture Notes in Computer Science, vol. 1389, pp. 390–410. Springer, Berlin Heidelberg New York (1998)
5. Chhabra, A.K.: Graphic symbol recognition: an overview. In: Tombre, K., Chhabra, A.K. (eds.): Graphics Recognition—Algorithms and Systems. Lecture Notes in Computer Science, vol. 1389, pp. 68–79. Springer, Berlin Heidelberg New York (1998)
6. Clark, A.F., Courtney, P.: Databases for performance characterization. In: Stiehl, H.H., Viergever, M.A., Vincken, K.L. (eds.) Performance Characterization in Computer Vision. Kluwer, Dordrecht (2000)
7. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models: Their training and application. Comput. Vis. Image Underst. **61**(1), 38–59 (1995)
8. Cordella, L.P., Vento, M.: Symbol recognition in documents: a collection of techniques? Int. J. Doc. Anal. Recognit. **3**(2), 73–88 (2000)
9. Courtney, P., Thacker, N.A.: Performance characterization in computer vision: the role of statistics in testing and design. In: Blanc-Talon, J., Popescu, D.C. (eds.) Imaging and Vision Systems: Theory, Assessment and Applications. NOVA Science, Hungtington, NY (2003)
10. Delalandre, M., Trupin, E., Ogier, J., Labiche, J.: Contextual system of symbol structural recognition based on an object-process methodology. Electron. Lett. Comput. Vis. Image Anal. **5**(2), 16–29 (2005)
11. Ghosh, D., Shivaprasad, A.P.: An analytic approach for generation of artificial hand-printed character database from given generative models. Pattern Recognit. **32**, 907–920 (1999)
12. Guyon, I., Haralick, R.M., Hull, J.J., Phipliops, I.T.: Data sets for OCR and document image understanding research. In: Bunke, H., Wang, P.S.P. (eds.) Handbook of Character Recognition and Document Image Analysis, pp. 779–800. World Scientific, Singapore (1997)
13. Haralick, R.: Performance characterization in image analysis: thinning, a case in point. Pattern Recognit. Lett. **13**, 5–12 (1992)
14. Hilaire, X.: A matching scheme to enhance performance evaluation of raster-to-vector conversion algorithms. In: Proceedings of 7th International Conference on Document Analysis and Recognition, vol. 1, pp. 629–633. Edinburgh, Scotland (2003)
15. Kanungo, T., Haralick, R.M., Baird, H.S., Stuetzle, W., Madigan, D.: Document degradation models: parameter estimation and model validation. In: Proceedings of IAPR Workshop on Machine Vision Applications, Kawasaki (Japan), pp. 552–557 (1994)
16. Kanungo, T., Haralick, R.M., Baird, H.S., Stuezle, W., Madigan, D.: A statistical, nonparametric methodology for document degradation model validation. IEEE Trans. Pattern Anal. Mach. Intell. **22**(11), 1209–1223 (2000)
17. Lladós, J., Valveny, E., Sánchez, G., Martí, E.: Symbol recognition: current advances and perspectives. In: Blostein, D., Kwon, Y.-B. (eds.) Graphics Recognition—Algorithms and Applications. Lecture Notes in Computer Science, vol. 2390, pp. 104–127. Springer, Berlin Heidelberg New York (2002)
18. Lopresti, D., Nagy, G.: Issues in ground-truthing graphic documents. In: Blostein, D., Kwon, Y.-B. (eds.) Graphics Recognition—Algorithms and Applications. Lecture Notes in Computer Science, vol. 2390, pp. 46–66. Springer, Berlin Heidelberg New York (2002)
19. Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R., Ashida, K., Nagai, H., Okamoto, M., Yamamoto, H., Miyao, H., Zhu, J., Ou, W., Wolf, C., Jolion, J.M., Todoran, L., Worring, M., Lin, X.: ICDAR 2003 robust reading competitions: entries, results, and future directions. Int. J. Doc. Anal. Recognit. **7**(2-3), 105–122 (2005)
20. Mariano, V.Y., Min, J., Park, J.-H., Kasturi, R., Mihalcik, D., Li, H., Doermann, D., Drayer, T.: Performance evaluation of object detection algorithms. In: Proceedings of the 16th International Conference on Pattern Recognition, Quebec (Canada), vol. 3, pp. 965–969 (2002)
21. Philips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The feret evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**(10), 1090–1104 (2000)
22. Phillips, I.T., Chhabra, A.K.: Empirical performance evaluation of graphics recognition systems. IEEE Trans. Pattern Anal. Mach. Intell. **21**(9), 849–870 (1999)
23. Tombre, K., Chhabra, A.K. (eds.): Graphics Recognition—Algorithms and Systems. Lecture Notes in Computer Science, vol. 1389. Springer, Berlin Heidelberg New York (1998)
24. Valveny, E., Dosch, Ph.: Performance evaluation of symbol recognition. In: Marinai, S., Dengel, A. (eds.) Document Analysis Systems VI – Proceedings of 6th IAPR International Workshop on Document Analysis Systems, Florence (Italy). Lecture Notes in Computer Science, vol. 3163, pp. 354–365. Springer, Berlin Heidelberg New York (2004)
25. Valveny, E., Dosch, Ph.: Symbol recognition contest: a synthesis. In: Selected Papers from 5th International Workshop on Graphics Recognition, GREC'03. Lecture Notes in Computer Science, vol. 3088, pp. 368–385. Springer, Berlin Heidelberg New York (2004)
26. Wenyin, L., Dori, D.: A protocol for performance evaluation of line detection algorithms. Mach. Vis. Appl. **9**, 240–250 (1997)
27. Wenyin, L., Zhai, J., Dori, D.: Extended summary of the arc segmentation contest. In: Blostein, D., Kwon, Y.B. (eds.) Graphics Recognition: Algorithms and Applications, Selected Papers from 4th International Workshop on Graphics Recognition, GREC'01. Lecture Notes in Computer Science, vol. 2390, pp. 343–349. Springer, Berlin Heidelberg New York (2002)
28. Wilson, C.L., Geist, J., Garris, M.D., Chellappa, R.: Design, integration and evaluation of form-based handprint and OCR systems. Technical report, National Institute of Standards and Technology, Technical Report NISTIR 5932 (1996)
29. Zhang, Y.J.: A survey on evaluation methods for image segmentation. Pattern Recognit. **29**(8), 1335–1346 (1996)