Report on the Second Symbol Recognition Contest

Philippe Dosch, Ernest Valveny

Report on the Second Symbol Recognition Contest

Philippe Dosch
Université Nancy 2
LORIA, UMR 7503
615, rue du jardin botanique,
B.P. 101, 54602
Villers-lès-Nancy Cedex, France
Philippe.Dosch@loria.fr

Ernest Valveny
Computer Vision Center
Computer Science Department
(UAB)
Edifici O, Campus UAB
08193 Bellaterra, Spain
ernest@cvc.uab.es

Abstract

Following the experience of the first edition of the international symbol recognition contest held during GREC'03 in Barcelona, a second edition has been organized during GREC'05. In this paper, first, we bring to mind the general principles of both contests before presenting more specifically the details of this last edition. In particular, we describe the dataset used in the contest, the methods that took part in it, and the analysis of the results obtained by the participants. We conclude with a synthesis of the contributions and lacks of these two editions, and some leads for the organization of a forthcoming contest.

1 Introduction

1.1 General principles of performance evaluation

For many areas within pattern recognition and graphics recognition, performance evaluation has become a crucial field of research work [1, 2, 3, 4]. This effort has become necessary in order to be able to compare different methods on standard datasets using metrics agreed by the research community. In general, all these evaluation works rely on several components:

- A *dataset* containing a sufficient number of representative data for the field under evaluation. Data can be either real or synthetic, depending on the application domain. It should also include several kinds and levels of degradation and deformation.
- A *ground-truth* that represents the perfect labelling of test data and therefore, the results that the participants are expected to provide.

- A *metric* to measure the distance between the ground-truth and the results provided by the participant methods.
- A *protocol* that specifies how the organizers and the participants exchange all information (input data, results, etc.) concerning the competition.
- A set of tools for the *analysis of results*. This analysis can be led from two different viewpoints: a data viewpoint, in order to determine how each kind of input data is recognized according to different methodological approaches, and a methodological viewpoint, in order to determine the strengths and weaknesses of every method for different kinds of data.

Some of these evaluation campaigns are designed to determine a sorting of the participant methods, based on a global performance measure computed after applying each method to the whole set of data. This approach is only possible in some domains where it is realistic to compute a global performance measure according to the characteristics of the data. However, whatever the performance measures are, we strongly believe that the main objective of an evaluation framework must be the scientific analysis of the results. This analysis must be intended to determine the different qualities expected for recognition methods: robustness, genericity, precision, computational efficiency. Usually, each of these qualities must be estimated with different performance measures computed over several sets of data.

These principles being defined, we would like to point out that complete and really useful performance evaluation requires a lot of tests, led under a large number of criteria. Usually, contests can only work with a limited dataset, which means that they can play an important role as relevant milestones in the evaluation process, but they must be completed with other efforts (like regular and large tests) to allow a good understanding of the recognition methods for a particular application domain.

1.2 Symbol recognition contests

For performance evaluation of symbol recognition, the general principles exposed above are obviously the same. Two evaluation events concerning symbol recognition have already been held before this edition. The first one was during the 15th International Conference on Pattern Pattern Recognition (ICPR'00) [5]. The symbol library for that contest consisted of 25 electrical symbols, which were scaled and degraded with a small amount of binary noise. Following this event, a second contest was held during the fifth International Workshop on Graphics Recognition (GREC'03) [6, 7], known as the first international contest on symbol recognition, as its characteristics were closer to those expected for such an event: several application domains, more symbols, more test images, different kinds and levels of degradation and noise, ... The contest organized in the context of GREC'05 and explained in this paper was the natural continuation of this one.

As there are many factors which can influence the performance of a symbol recognition method, the main goal of these contests were not to give a single performance measure for each method, but to provide a tool to compare various symbol recognition methods under several different criteria. From an evaluation viewpoint, the question consists of determining the performance of symbol recognition methods when working on various kinds of symbols, extracted from diverse application domains, under several constraints, with different levels of noise and degradation.

In the following of this paper, first in section 2, we will briefly review the main features of the first edition of the contest during GREC'03. Then, in section 3 we will explain the second edition of the contest: the differences with the first edition, the dataset and the development of the contest. In section 4 we will give some details about the methods that took part in it and section 5 is devoted to the analysis of their results. As we have explained before, evaluation effort should not be limited to some specific milestones but it should have a continuity over time. In section 6 we will explain the main goals of the project Épeires, a project currently under development intended to provide a stable framework for evaluation of symbol recognition. Finally, in section 7 we draw the main conclusions of the contest and some hints about future work.

2 First edition of the contest

To have a complete overview of the first edition of the contest, we advise to refer to [6]. Here, we will only remind the main features of this first edition:

- 50 different model symbols were used in the contest, from two different application domains (electronics and architecture) and composed of lines and arcs.
- Only segmented images of symbols were used.
- Scalability: 3 sets of symbols each with 5, 20 and 50 symbols were defined to test the robustness to scalability.
- Generation of images with rotation and scaling in order to test the invariance to geometric transformations.
- Generation of images with 9 models of binary degradations, following the model of degradation of Kanungo [8], see figure 1.
- Generation of images with 3 models of vectorial shape deformations, see figure 2.
- Generation of tests with combination of these transformations.
- Whenever possible, pixel-based and vector-based test images were provided to participants.
- 5 participants took part in the contest.

- More than 7000 test images were synthetically generated and organized in more than 70 independent tests.

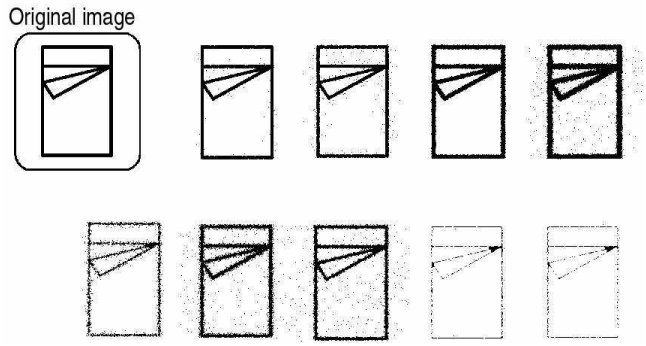


Figure 1: The nine models of degradation used for the first edition of the contest.

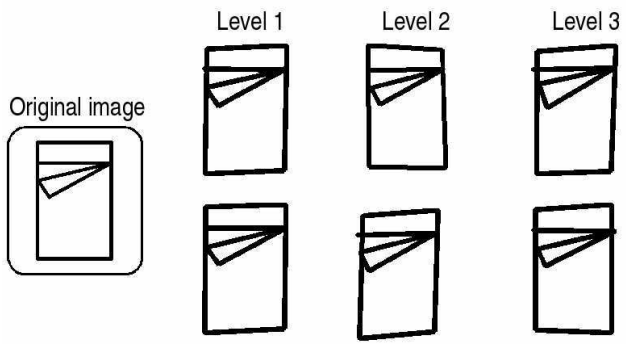


Figure 2: The three models of deformation used for the first edition of the contest.

The analysis of results was led according to the scientific criteria exposed above, i.e, trying to study how the performance of methods was degraded when applying transformations and noise to the original images. The main conclusions were:

- With respect to the different tests supplied, most of participant methods had very good recognition rates, usually between 95% and 100%, when dealing with 5 or 20 model symbols and few degradation and deformation.
- In general, the performance decreases with the number of symbols, for all kinds of tests.
- Methods are not fully invariant to rotation and scaling, even whenever they claim the contrary.

- Robustness to degradation: performance degrades significantly with heavy noise or when connectivity of lines is lost.
- Robustness to vectorial shape deformation: The performance decreases with the deformation, but not too much.

As a conclusion of the first edition, and as the results were quite good, we planned to increase the complexity of the data used in a future contest in these directions:

- Add more symbol models, in order to evaluate the scalability of the recognition methods.
- Add new models of noise (heavy noise), in order to evaluate more accurately their robustness.
- Define tests with non-segmented symbols, in order to evaluate the ability to localize, segment and recognize these symbols in real drawings.

3 Second edition of the contest

3.1 General principles

Following the conclusions of the first edition of the contest, we have tried to set up a new edition including the new features pointed out at the end of previous section. We have been able to achieve only two of those goals. In this new edition, we have included more symbols and more models of noise. However, we have not succeeded in the inclusion of non-segmented symbols. In fact, a lot of effort is required in order to set up the evaluation of the localization and segmentation of symbols. Among the issues to be addressed we can remark the following:

- To build a dataset providing a large and enough number of real images of different domains, such as architecture drawings, electronic maps, etc. As these data are often private, it is difficult to get a dataset representative enough for such a contest.
- The definition of metrics allowing the comparison of ground-truth with the results provided by the participants. The ground-truthing itself requires a lot of time, and has to respect a very well defined methodology to be fully exploitable. In particular, we believe that the definition of the ground-truth have to include the creation itself, but also the validation by different people, in order to ensure that the ground-truth will be agreed by everyone. Incidentally, ground-truthing is a very time-consuming task, as test data have to be handled by many people to become fully exploitable.
- The design of an environment allowing the automatic processing of the results provided by participants, in order to analyze them. Whereas the

evaluation of symbol recognition methods requires a reasonable framework of evaluation, the evaluation of localization and segmentation methods requires a significant bigger effort as much data have to be managed.

In this context, we are working in parallel on a project, funded by the French government, which aims at providing such an environment for the scientific community. This project, called ÉPEIRES, is briefly presented in section 6. Once this project has been fully developed, it will allow to easily define tests for symbol localization available to everybody.

In conclusion, for this edition of the contest, we have given up the idea of including non-segmented images and we have only defined some tests with segmented images, very similar to those proposed during the first edition, but with the following remarkable differences:

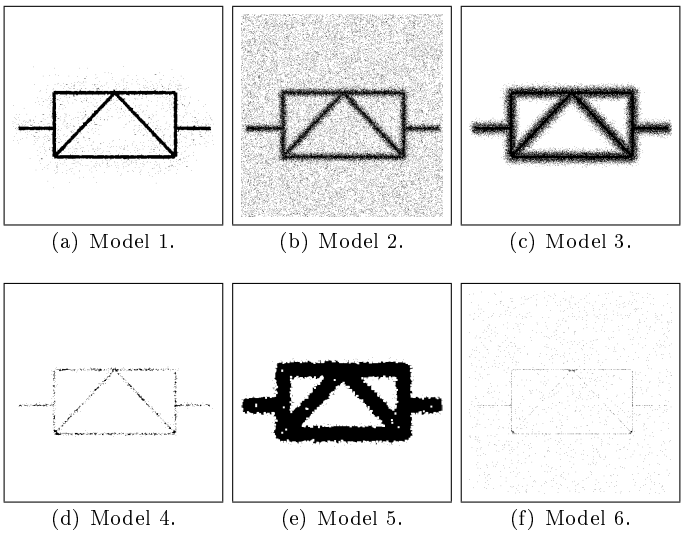
- The set of symbols has grown from 50 to 150 different symbols, allowing the definition of tests useful for the evaluation of the scalability of recognition methods.
- Four new degradation models have been added, allowing the generation of more noisy synthetic data. These degradation models are further explained in the next section.
- Tests have only included bitmap images for this edition. Vectorial images have not been taken into account as the selected models of degradation do not allow for a good vectorization of images.

3.2 The symbol database

As for the first edition, two application domains have been mainly used, architecture and electronics. We have used 150 different symbols, some of them with similar shapes, grouped in four sets containing respectively 25, 50, 100 and 150 symbols, in order to evaluate the scalability of recognition methods. As we have previously said, we have only considered in this edition presegmented symbols, *i.e.* images containing one instance of one symbol, and only bitmap format.

Several transformations, global transformation and noise, have been applied on these ideal models, in order to evaluate the robustness of the recognition methods to such transformations. Global transformations include rotated and scaled images, whereas noisy images have been generated using the well-known Kanungo's method [8]. We remind that the initial purpose of this method is to modelize the noise produced by operations like printing, photocopying, or scanning. The method is formal and validated for its correctness, but the determination of the set of parameters used for the contest is more empirical.

In the first edition of the contest, we have tried to reproduce a set of degradations reproducing some realistic artifacts as those mentioned above. But for this edition, six degradation models have been defined, aiming at constituting a set of what we could call "torture models" rather than some realistic degradation models, as this issue was relatively well addressed during the first edition.



(a) Model 1. (b) Model 2. (c) Model 3.

(d) Model 4. (e) Model 5. (f) Model 6.

Figure 3: The six models of degradation used for the second edition of the contest.

This way, we can test the robustness of recognition models under very extreme conditions. Some samples of images generated with these models are shown in figure 3.

But at the same time, we have to be very careful on the conclusions we can draw on the related results. We are obviously aware that it may be dangerous to rely the performance evaluation on some too noisy synthetic data, probably too different of real images. So, if this dataset can be used to proof the robustness of the participant methods under extreme conditions, we have also to be careful on the meaning of the evaluation, especially on the capacity of these methods to work on real data.

3.3 The contest

All information related to the second edition of the contest is available at `http://symbcontestgrec05.loria.fr/`. Most of this information, especially those points related to the formats and protocols, is the same as for the first edition. The report of the first edition [6] and the related Web site at `http://www.cvc.uab.es/grec2003/SymRecContest/index.htm` provide a good description of the contest environment.

As for the first edition, independent tests have been designed with respect to different categories (concerning the number of symbols, the kind of noise, etc.) so that each participant, according to the specificities of its method, could choose the tests he wanted to run. As some methods require training data in order to work properly during the contest, the models of all the symbols and some sample tests were made available for all participants before the contest,

with the associated ground-truth. The tests provided to the participants in the contest were similar, but different from the sample tests.

For the second edition, tests have been designed according to the following categories:

- *Scalability*, with 4 categories of tests involving an increasing number of symbols: 25, 50, 100 and 150. This category is intended to evaluate the capacity of the recognition method to discriminate symbols as the number of models increases.
- *Degradation models*, with the 6 models presented in section 3.2. This category is intended to evaluate the robustness of the methods when symbols are degraded under several conditions.
- *Transformations*, by considering rotation and scaling, either alone or together. In addition, a category without any transformation has been defined. As for degradation models, this category is intended to evaluate the robustness of recognition methods under geometric transformations.

All these categories have been combined resulting in 96 different tests, with a total number of 6000 test images.

4 Participants

In this edition, four participants and their method took part in the contest. Two of these participants have a paper describing their method in the current LNCS volume. In this section, we only give a brief overview of the most relevant features of the participant methods, as provided by their authors.

4.1 Jing Zhang, City University, Hong Kong

The recognition method is a statistical, pixel-based method. The method used is very similar to Su Yang's. The symbol descriptor we used is referred to as Structural Feature Histogram Matrix (SFHM), which is an improvement of Yang's SIHA in two aspects:

1. SFHM computes length ratios and angles via a symbol's centroid;
2. SFHM integrates the information of length ratios and angles.

4.2 Min Feng, City University, Hong Kong

The recognition method is a statistical and pixel-based method. The similarity is calculated by matching the point sets extracted from the symbols. The assumption in the method is many to many correspondence, which reduces the time complexity into O(n2). However, the new similarity function is not invariant to rotation. To recognize rotated symbols, we compute their angular

distributions and align them by their orientations. The whole recognition procedure consists of three steps: image compression, denoising and recognition. Firstly, the input images are compressed in order to cut down the number of foreground points. After compressed, each pixel indicates the density of the foreground points in the original image. Secondly, a novel denoising technique is utilized to remove the noises from the compressed images. Finally, the above similarity function is used to compute the similarity between each pair of preprocessed test symbol and model symbol, and then for each test symbol the best matched symbol model is outputted.

4.3 Wan Zhang, City University, Hong Kong

The method is a statistical approach, where a symbol is represented by a 2D joint density estimated from a set of points sampled from the skeleton of the symbol. Matching two symbols is then equivalent to determining whether the two symbols have a similar probability distribution or not. In other words, if the points on the test symbol fit the density of the symbol model well, we can determine that the test symbol is similar to the model one. By adopting the Kullback-Leibler (KL) divergence as a distance of the two distribution densities, the similarity of the two symbols can be measured. In the first preprocessing module, a freeware (Ras2Vec) is selected to finish the vectorization processing of the binary images and obtain the skeletons of images. Furthermore if necessary, a few preprocessing techniques will be applied to reduce the noise and help to improve the robustness. The method is independent of the position of the symbol, and easy to be extended for rotation-invariance and scale-invariance.

4.4 Andyardja Weliamto, Nanyang Technological University, Singapore

This recognition system is based on the statistical approach. It assumes that at the end of the preprocessing step we have single pixel thin line. The system consists of several steps:

1. Preprocessing/filtering: adaptive noise preprocessing using morphological, convolution and thresholding for different noise models (noise model classification).

2. Feature selection/feature vector composition based on Fisher Discriminant Analysis.

3. Classification based on the k-Nearest Neighbor with Mahalanobis distance.

The problem of the system in the contest was that we did not have enough time to verify the linearity of preprocessing image among different noise models. We also needed to test some parameters that deal with the training system. The constraints of the system are: first, it is difficult to make the preprocessing of the image linear among different noise models without experimentally testing

and second, the feature vector should be unique with higher discriminant factor. That is why the recognition rate drop since the preprocessing step fails. Incorporating better adaptive noise reduction preprocessing of images increases the recognition rate of the system by 10%. Another problem was that some symbols have similar radial feature. Therefore, we need to introduce some new features based on the angular feature and as a result of it the recognition rate increases by 6%.

5 Analysis of results

5.1 Introduction

As a preamble of this section, we want to point out that this analysis is related to the dataset defined, which contains only synthetic data, degraded using some set of parameters for Kanungo's method, as explained in section 3.2. Even if some of the generated data seems close to real data, as those represented in various technical documents, other images are rather far from real or realistic data. The purpose of this kind of contest is obviously to determine what methods work on real data, as this is the typical way they are used in real applications. But as building a set of real data, with a representative and sufficient number of images, is complex for several reasons (availability, rights, work force), the current edition is partially based on exaggerated noisy data. This is a more practical way to proof the robustness of recognition methods, but it also implies that we have to keep aware of these evaluation conditions, and therefore, be careful on the interpretation of results.

Moreover, in this edition, some of the tests have been designed with a low number of images for some categories, as the participants had to run their method on all tests the day before GREC. It leads sometimes to strange recognition rates, as tests did not contain a significative enough number of images, and maybe some images were "easier" to recognize in some of the sets.

These two constraints, the use of synthetic data and the restricted number of images involved, lead of course to some limitations for the determination of the more generic and robust method/approach in symbol recognition.

Another important remark we also want to point out is that participant methods should not integrate *a priori* knowledge about degradation or transformation models, if the goal has to be the evaluation of the genericity. From a scientific viewpoint, the most important task is to evaluate the core of symbol recognition methods, and not frameworks integrating pre or post processing steps dedicated to the degradation models. Even if some labels are provided for each proposed test, either training or final, they are essentially tips to easily determine if a recognition method is adapted to a test, not to allow an adaptation of the method to the test. As this principle was not explicitly defined for this edition, it may be another limitation, as some methods included specific preprocessing depending on the kind of noise detected in the test.

In the following, we will discuss the results provided by the participants,

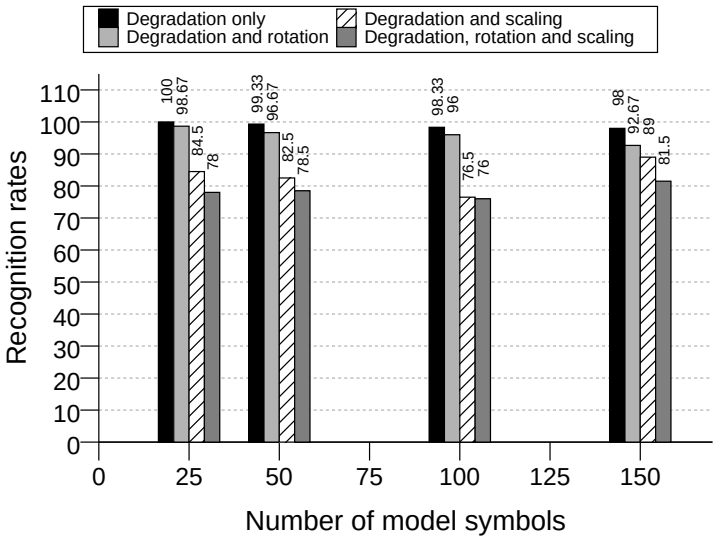


Figure 4: Synthetic chart showing the average recognition rate obtained by all participant methods for the first degradation model with all combinations of rotation and scaling.

from several viewpoints, taking into account the above stated limitations and bearing in mind that our main goal is not to give an absolute winner, but to show the robustness of the methods to the different evaluation criteria.

5.2 Clean images

The first degradation model was designed to simulate some clean images, *a priori* very close from real ones. The results obtained for the tests using this model are presented in figure 4 (see the first column). Recognition rates correspond to the average rate for all participant methods. For comparison, figure 5 shows the results obtained by the best participant method, those of Feng Min, with an average recognition rate of 94.88% (see the first column too). The results are very good, for all methods. Even if the recognition rate decreases a bit when the number of symbols increases, the average recognition rate is equal to 98% when dealing with 150 different symbols. So we can conclude that symbol recognition is quite mature with these contest conditions, close to ideal real ones. The next step in order to evaluate the scalability of the methods under these conditions is probably to propose tests with a very larger number of symbols, maybe 1000.

5.3 Clean images with transformations

Still working with the first degradation model, close in our opinion to ideal real images, tests have been defined with rotation, scaling, and a combination of

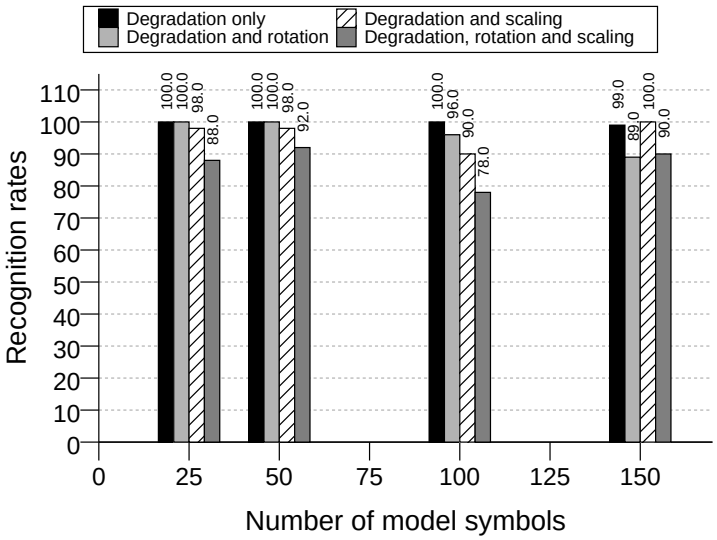


Figure 5: Results obtained by the best participant method, those of Feng Min, for the first degradation model.

both transformations. The corresponding results are also presented on figure 4 and 5 (second, third and fourth columns). When dealing only with rotation, the recognition rate decreases and this tendency is accentuated when the number of symbols increases. The average recognition ability is almost reduced by approximately 5% when dealing with 150 symbols with respect to the same tests without any transformation. Similar remarks can be done with the tests related to scaling only and the combination of both transformations, leading respectively to approximate reductions of 9% and 16%.

The tendencies shown on the synthetic chart (figure 4), presenting only average recognition rates, are similar to results obtained by each participant. For Feng Min (figure 5), we can however see that the results obtained with the test dealing with scaling only and 150 model symbols are better than the others related to this scalability category. As this test contains only 50 test images, which is probably not representative enough with respect to the 150 model symbols, it is however difficult to formally interpret this result.

But from a general viewpoint, transformations clearly still impact recognition quality.

5.4 Scalability

Testing scalability with respect to the number of considered model symbols is one of the main objectives of the symbol recognition contests. For this edition, figure 6 presents a synthetic chart of the average results obtained for all degra-

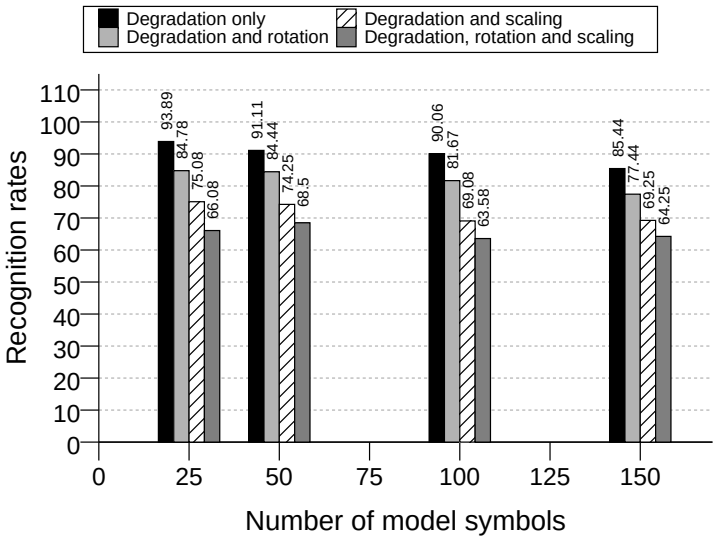


Figure 6: Synthetic chart for all degradation models.

dation models, according to the number of model symbols. The performance clearly decrease when the number of symbols increases, with some variations according to the kind of tests. For tests including degraded images without any transformation, the loss of recognition is about 8.5% when the number of symbols evolves from 25 to 150. This loss decreases when transformations are added. It is about 7.3% when rotation is added, about 5.8% when scaling is added and only about 1.8% when both of these transformations are added. This is a bit surprising, as we intuitively expect that the more constraints are added, the more performance decreases.

In general, the decrease seems to be linear with respect to the number of symbols involved. A larger number of model symbols has to be considered in further events related to performance evaluation to allow a more detailed analysis of scalability impact.

5.5 Participants method and degradation models

The last chart, presented in figure 7, shows the average recognition rates obtained by each participant for each degradation model. The recognition rates obtained for each degradation model is rather different from one participant to another. This fact shows that, according to the recognition approach, methods are more or less sensitive to the kind of degradation. It reminds the importance of using several degradation models for this kind of performance evaluation. It would be interesting to have more details on each participant method to have a better understanding of this behavior. The model 6 appears to be the more

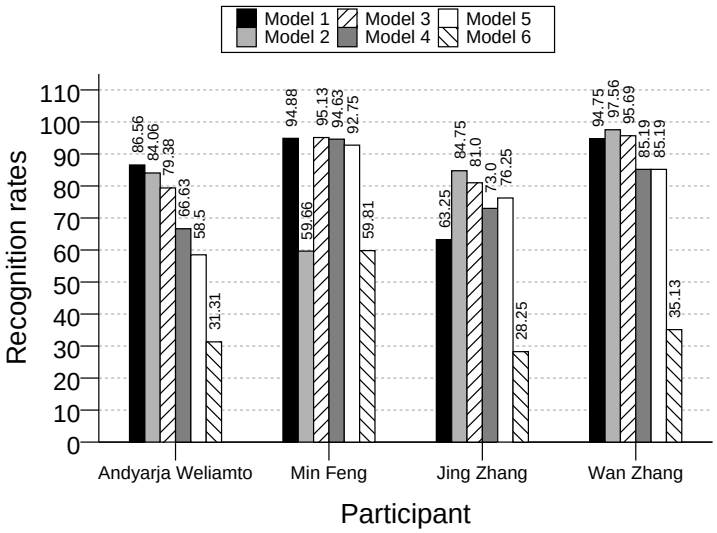


Figure 7: Synthetic chart for each participant with respect to the degradation models.

Table 1: Overall results for each participant.

Participant	**Average recognition rate**
Andyardja Weliamto	70.28%
Min Feng	83.33%
Jing Zhang	67.65%
Wan Zhang	82.82%

difficult to recognize in general. This is probably because the corresponding images are very degraded, with loss of connectivity and a global noise with pixel density close to that of the symbol itself. The only method having a recognition rate higher than 50% for this model is that of Min Feng, with a recognition rate equal to 59.81%.

Min Feng is also the global winner of the contest, with an average recognition rate of 83.33% for all the proposed tests, as shown in table 1. However, as previously stated, the main purpose of this contest is not to determine a winner, but rather to study the evolution of the recognition rates according to the test characteristics. But it appears that 83.33% is a good overall recognition rate considering the proposed tests, most of them designed to be "torture tests", *i.e.* exaggerated noisy data proposed to assess the robustness of the participant methods.

5.6 Synthesis

Following the experience of the first edition of the symbol recognition contest, this second edition has been organized in order to propose some difficult tests on segmented symbols. It appears that the recognition rates are quite good with respect to some of the degradation models proposed. In particular, in conditions close to the ideal ones, an average recognition rate of 98% has been reached by the participant methods. As a conclusion, one can say that symbol recognition, in the conditions defined for the contest, is quite mature, even if participants methods are not fully invariant to transformations like rotation and scaling.

As expected, performances generally decrease when the number of model symbols increase and when transformations are added. Therefore, more models have to be proposed to accurately measure the scalability with really large sets of symbols. And more data, representative of other application domains, have to be supplied too, in order to evaluate the robustness of the participant methods to different domains and kinds of symbols.

Now, we think that the next important challenge is to organize tests about symbol localization, that is to say, symbol recognition on images including several instances of different symbols in their real context, connected to other lines or elements of a drawing. Tests about symbol recognition are still interesting, but only in some particular aspects, such as, for example, scalability with a large number of model symbols.

We plan to define forthcoming tests about symbol localization thanks to the Épeires project presented in the next section.

6 The Épeires Project

The Épeires Project[1] is funded by the French Ministry of Research in the context of the Techno-Vision Campaign[2]. Its purpose is the construction of a complete environment providing tools and resources for performance evaluation of symbol recognition and localization. The aim is to estimate their capacity to recognize and to localize symbols in a generic way, according to various criteria: application domain, modelization, number of symbols involved, document quality, etc. The consortium is currently composed by 6 laboratories (City University of Hong Kong, CVC Barcelona/Spain, LI Tours/France, L3I La Rochelle/France, LORIA Nancy/France and PSI Rouen/France) and also 2 French companies (France Télécom R&D and Algo'Tech Informatique). This environment is intended to be used by the whole scientific community.

Document analysis generally deals with two main kinds of symbols: structured symbols and logos. In the Épeires Project, we intend to consider symbol recognition as a whole, without making any particular distinction between them. Participant methods should be subsequently tested on both kinds of symbols. The Épeires Project is organized along 3 main directions:

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

[2] http://www.recherche.gouv.fr/appel/2004/technovision.htm

- *Development of a database of test images*, in order to get a large variety and a large number of test data, possibly free of rights. Images will be proposed in clean and degraded versions, to test the robustness of the recognition methods. A ground-truth will be associated with each image, using a collaborative software (currently under development) connected to the information system of the project.

- *Design of metrics and protocols* specifying how the results will be analyzed.

- *Performance evaluation of the methods supplied by the participants.* It will determine the methods providing the best recognition and/or localization rates on the documents of the test database. It will also be the opportunity to measure the strengths and weaknesses of the methods. As for the contests, the goal will not only be to determine the most reliable chains of applications from a synthetic viewpoint, but also to understand the influence of the different approaches on the quality of the results.

As a result, it is planned to provide at least 1000 model symbols and 100000 test images. We hope we will provide to the community a great tool for performance evaluation of symbol recognition and localization.

7 Conclusion and next steps

As a conclusion of the two contests on symbol recognition organized during GREC'03 and GREC'05, we would like to point out the following issues:

- More information is needed from the participants to better understand the recognition rates. We expect that they give a more detailed description of their methods, and they give more feedback on their results. We plan to provide tools to assist these descriptions and discussions.

- We have to provide facilities allowing to spread and analyze the results of evaluation campaigns, for the further contests as well as for any campaign related to symbol recognition and localization. We hope that the ÉPEIRES project will supply such facilities very soon.

- More data, free of use, are still required, as performance evaluation cannot be fully suitable without a large number of heterogeneous data. It is a call for the community, as we all need these data to make evaluations on our methods.

- No new degradation models based on Kanungo method are needed. After these two contests, we have defined 15 different models, from more realistic noise to "torture models", and we think it is enough. It is more interesting to support new kind of noises, like scratches, or to mix the existing models in blind tests.

- Next campaigns must include blind tests in order to ensure that participant methods are not adapted to the particular data of the contest. We would like to be sure that participants address the good goal: design generic symbol recognition methods, working with all kind of (noisy) symbols, and not only those provided in the context of these contests.
- Campaigns of evaluation must be led more regularly than every 2 years. If we fully want to integrate performance evaluation as a main part of each research on symbol recognition method, we need a stable environment for evaluation events with more heterogeneous data.
- And of course symbol localization must be addressed as it is currently on of the main challenging issues for the symbol recognition community.

For the major part of these remarks, we hope that the ÉPEIRES project will be able to provide such a complete framework to the community.

Acknowledgment

The authors would like acknowledge the participants for their participation to the contest and for their contribution to this article. They also would like to acknowledge the French Ministry of Research for the funding of the ÉPEIRES project as a part of the Techno-Vision campaign. This work has also been partially supported by the Spanish project CICYT TIC2003-09291.

References

[1] Kong, B., Phillips, I.T., Haralick, R.M., Prasad, A., Kasturi, R.: A benchmark: Performance evaluation of dashed-line detection algorithms. In Kasturi, R., Tombre, K., eds.: Graphics Recognition: Methods and Applications, Selected Papers from First International Workshop on Graphics Recognition, GREC'95. Springer, Berlin (1996) 270–285 Volume 1072 of Lecture Notes in Computer Science.

[2] Chhabra, A., Phillips, I.: The second international graphics recognition contest - raster to vector conversion: A report. In Tombre, K., Chhabra, A., eds.: Graphics Recognition: Algorithms and Systems, Selected Papers from Second International Workshop on Graphics Recognition, GREC'97. Springer, Berlin (1998) 390–410 Volume 1389 of Lecture Notes in Computer Science.

[3] Chhabra, A., Philips, I.: Performance evaluation of line drawing recognition systems. In: Proceedings of 15th. International Conference on Pattern Recognition. Volume 4. (2000) 864–869 Barcelona, Spain.

[4] Wenyin, L., Zhai, J., Dori, D.: Extended summary of the arc segmentation contest. In Blostein, D., Kwon, Y., eds.: Graphics Recognition: Algorithms

and Applications, Selected Papers from Fourth International Workshop on Graphics Recognition, GREC'01. Springer, Berlin (2002) 343–349 Volume 2390 of Lecture Notes in Computer Science.

[5] 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. Volume 4. (2000) 870–876 Barcelona, Spain.

[6] Valveny, E., Dosch, P.: Symbol recognition contest: a synthesis. In Lladós, J., Kwon, Y.B., eds.: Graphics Recognition: Recent Advances and Perspectives – Selected papers from GREC'03. Volume 3088 of Lecture Notes in Computer Science. Springer-Verlag (2004) 368–385

[7] Valveny, E., Dosch, P.: Performance Evaluation of Symbol Recognition. In: Proceedings of the 6th IAPR International Workshop on Document Analysis Systems, Florence, (Italy). Volume 3163 of Lecture Notes in Computer Science. (2004) 354–365

[8] 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). (1994) 552–557