This document is a discussion report, not a regular paper.

Performance Evaluation of Symbol Recognition/Spotting Systems: A Report of Discussions

Ophelia Ezra[1]

April 8, 2008

Abstract

How to evaluate the symbol spotting/recognition? This document reports discussions done by a working group on this topic between December 2007 and February 2008. This work group was composed of the following people: Alicia Fornes (CVC), Dimosthenis Karatzas (CVC), Ernest Valveny (CVC), Hervé Locteau (LITIS), Jean-Pierre Salmon (LORIA), Jean-Yves Ramel (LI), Marçal Rusinol (CVC), Mathieu Delalandre (CVC), Philippe Dosch (LORIA), Rashid Qureshi (LI) and Tony Pridmore (SCSIT). The corresponding institutes are the CVC (Barcelona, Spain), the LORIA (Nancy, France), the L3i (La Rochelle, France), the LI (Tours, France), the LITIS (Rouen, France) and the SCSIT (Nottingham, UK). This document reports also some comments done during the EPEIRES Meetings of May 2005 [15] and January 2006 [16].

Content

1. Introduction

1.1 Evaluation purposes

Mathieu: The first things we have discuss concerns the purposes of the Contest. These ones will depend obviously of the systems' applications. Based on my literature review they are four main applications: the symbol localization (i.e. segmentation) [20] [21], the symbol recognition [1], the symbol spotting [2] and the symbol mining [3].

- **Localization:** In this case the systems are interested only to locate the symbols especially to segment them (i.e. to crop the symbols from the localization data) or to perform a background/foreground separation [20] [21].
- **Recognition:** It is the usual one [1], the systems have to provide labels and localizations of symbols from entry test documents using a learning database.
- **Spotting:** Concerning the spotting [2] the systems have to provide ranked lists of localization (i.e. image path with a geometric object “point, box, contour, etc.”) from a query image (cropped by a user from a real drawing). We talk about QBE (Query By Example, see this Wikipedia page).
- **Mining:** Concerning the mining [3] the systems have to provide localization data and cluster labels of symbols (cluster1, cluster2, etc.). So it is same than a recognition process but without any previous knowledge concerning the models. I suppose that interests people with this application is more the indexing results of whole documents than the ones of symbol mining (the symbol mining is more a way to index the drawing than an objective). But in our works I think that an evaluation at the document indexing level is out of order.

Philippe: At this time the final purpose of the evaluation is the recognition. Spotting and mining are just use-cases we can adapt next.

Ernest: The evaluation system must work in two modes: spotting and recognition.

Mathieu: I'm not sure that all the people will be interested in testing the recognition. Most of the expected participants for the Contest seem working at a spotting level.

Philippe: It is not true. Some EPEIRES people work on recognition. Moreover, there are some ambiguities concerning the definition of “spotting”, that we call spotting can be considered as recognition. You can adapt a spotting method to do recognition by using learning database. A spotting method is particular adaptation of a recognition one.

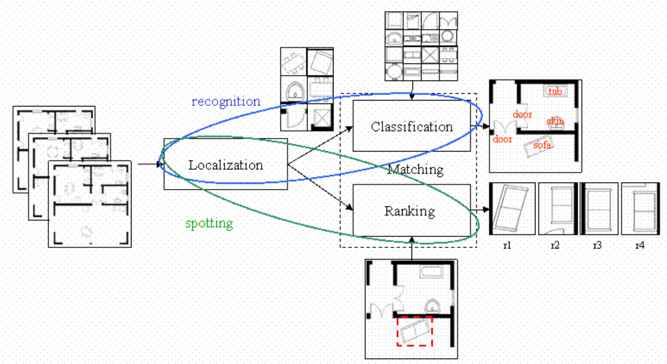
[1] Colective name generated with behindthename

Dimosthenis: Spotting and recognition are two different processes in the objectives. The first one has to find similar objects from a given query and the second one raises on “a priori” knowledge to recognize high semantic objects. That you have to do is to propose your own definitions of spotting and recognition for the Contest, and the purpose of the evaluation for these two applications.

Ernest: I think we're talking about the same thing. Recognition and spotting are very near processes, just two different applications of a pattern analysis system.

Ernest, Philippe: One definition is based on the use of previous knowledge. Recognition could have a learning step whereas the spotting has to run only with the QBE.

Mathieu: I propose here a figure detailing the links between the spotting and the recognition. Based on our discussions my felling is that only the matching level and the data to provide are different: classification (with a model database) for the recognition and ranking (with QBE) for the spotting. In both cases the processes relay on a localization/segmentation process.



Philippe, Jean-Yves: This means that the spotting systems have to retrieve using the full document database.

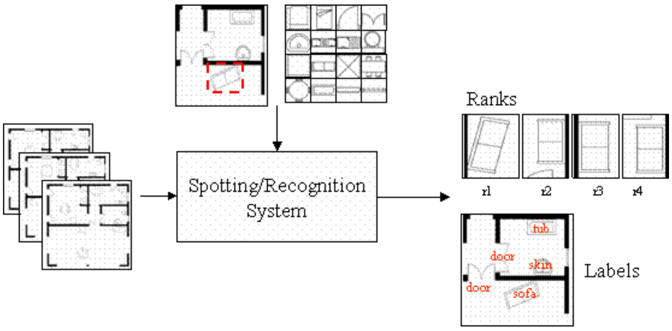
Marçal: Off course, it is the definition of the spotting.

Jean-Yves: I agree with the definitions of Ernest, Philippe, Marçal and the figure proposed by Mathieu. One thing, it seems to me that until now the recognition Contest corresponds only to the classification step.

Mathieu: Yes, exactly. I'm not sure if the steps' names are ok, it is just a proposal. We can also call the “classification” step “recognition”, in this case have you other proposals to label the “localization & recognition steps”: understanding, re-engineering, segmentation & recognition, etc.

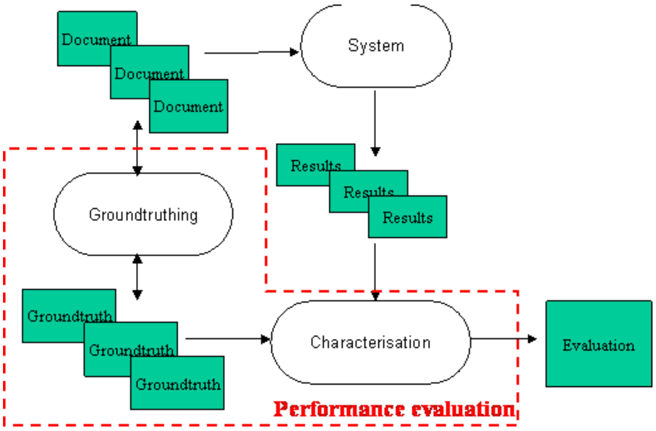
Marçal: In the real-life it is complicated to split the systems in tow parts. Some of them are black boxes in which the localization and the matching are done in a single step.

Mathieu: Ok, I propose here another figure.



Jean-Yves: The first one is better.

Mathieu: Also we have to define exactly what performance evaluation is. At this time I think we agree there are two main steps: the groundtruthing and the performance characterization.



Philippe: I agree with all these discussions and definitions. I suggest here to make priorities in our work and to focus first on the recognition (we will see then for the spotting).

1.2 Evaluation campaign

Mathieu: Concerning the evaluation campaign a first important point concerns the organization of the contest. Based on some discussion returns I think there is a dilemma between two ways of organization: day-contest vs on-line. The first one is the past one: the participants meet all together in a room the contest's day (or a range of day) to perform the tests. The second will be done in an on-line way. In this case the participants have to download the tests from a website, to perform them from their laboratories, and to upload their results. This will allow like this to do the Contests using largest time slots (several weeks or months). This second way is particularly adapted for the evaluations that require to compute large amount of data (like the spotting/recognition from whole drawings).

Philippe: At this time we're thinking to change the next organization of the Contest due to the spotting/recognition topic. It seems not reasonable to do this evaluation in the same way than the one on the segmented symbols: there is a complexity problem due to the use of whole drawings. Next Contests could be composed of two parts, a day-contest and an on-line one for the whole drawings.

Ernest: Yes but doing on-line tests will raise the problem of “cheaters”. The on-line solution has to be study carefully.

Philippe: Yes I'm ok, but I think we have to investigate this way. The performance evaluation side should apply a “confidence principle” with the participants. Every body will win in the collaboration between the recognition and evaluation people.

Ernest, Mathieu: We agree, the “cheaters” have basically no place in the research community. But the performance evaluation people have to check and to control the competition. This is waited for us by the community, it is our task. So we have to secure the competition and the day-contests are a good way to do it.

Mathieu: To answer a first thing is to determine if it is still possible to do day-contests or not.

Philippe: If we do day-contests it is then important to have images of weak size. Having images of small dimension allows to reduce the processing times of systems and then to respect the time constraints of the organization. However, what we will do of the whole images we have groundtruthed?

Jean-Yves: Ok, that you have to do is to select the small images and to reduce the number of image per test, not to reduce the images' resolution. We have to be realistic with the tests, not to change the tests due to the time constraints of the Contest organization.

Philippe, Ernest and Mathieu: We agree. In this case we have to deal with images of weak content (e.g. floorplans with a small number of room, diagrams of some simple functionalities, etc.), and propose tests composed of a few number of image.

Hervé: Ok, but a problem still remains. We're talking here always about the complexity of spotting/recognition systems, but what about the characterization ones? I think the characterization could raise also complexity problems (especially the localization evaluation from whole drawings). May be it will make impossible to continue the day-contests.

Mathieu: Yes, if we consider the last organizations of Contests (Contest the staring day of the Workshop and presentation of results at the end) we can rely on 24 - 72 h of computation. We have to take care to not overflow this limit.

Hervé: Yes, but we have lot of thing to evaluate (scalability, impact of the learning database size, noise level of images, number of QBE ....). In any cases we will still have complexity problems to evaluate all these aspects. To do the day-contests will be still restrictive to for the evaluation purposes.

Philippe: May be it exist a third way of organization that will solve all the problems. We can do the contest in a benchmark way. In this case we will ask to the participants to upload their systems (with a manual). We will run and evaluate next these systems our self (i.e. from the performance evaluation side).

Mathieu: I think it is a very interesting idea, but I'm a little bit afraid because it should take time and money to do it (i.e. to implement the online benchmark platform). Moreover, if we do it we have to define some copyright agreements with the participants (that is not so easy). Not all of them will be ok to give their systems without some previous guaranties.

Philippe: Yes, if we need several months to develop such benchmark system, only to secure the process for some participant it is then not a "good deal". We have already a lot to do in this field, not to waste time.

Hervé: From my point of view we can do a fourth way of organization: the grid-contests. In this case we have to define the minimum requirements for the day-contest, and next to make all the other ones optional, and to propose to do it in an on-line way.

Mathieu: Yes, we will be able next to compare the day vs. online contests: it will check any cheat problem. However, we must keep in mind that finding participants for a Contest is a harder task (see [4]). So it is necessary to allow every people to participate, not to discourage them. We will have to find a good threshold of what we cannot ask to the participants. May be to make free the participants to choose their own tests propose could avoid the renunciation cases.

Philippe: I agree, we have to allow all the "categories" of system to compete i.e. the most trained ones but also the "youngest". It is the reason for which we must still propose easy tests in the future (i.e. segmented symbol images with low distortion). We will combine these tests with harder ones, full drawing with high distortion levels.

Jean-Yves, Mathieu: There is also one thing we must to keep in mind is the "Contest effect". We mean by organizing Contests we format the researches done in the symbol recognition field. Contest databases become standard ones on which people have to test and to train their systems in order to publish and to be recognized in the community. However, Contest databases are not our "ultimate" goal. Many research applications have to be considered, and a system could be the best for a specific application and to have bad results on the Contest data. To solve this problem it is important to allow people to appropriate themselves the evaluation frameworks. It will allow them to constitute their own tests and to do their own evaluation. **Philippe:** It is already done in the EPEIRES platform. Tests built by a user could be uploaded on the server and to make public or private. However we have still to work on the reporting of results: how to upload and to store the results on the sever and to make them public or private.

2. Groundtruthing

2.1 Real-life approach

Mathieu: Concerning the groundtruthing from real-life document the first thing we have to do is to develop a document web manager. It will allow to collect document images and to add corresponding metadata. At this time there is no document collection allowing to perform experimental tests (I mean here documents we can use now without the groundtruth

Philippe: It is done in the EPEIRES platform.

Mathieu: How does work the groundtruthing tool of the EPEIRES platform.

Epeires report 2006: The groundtruthing tool from real life documents has to be the more ergonomic as possible. The retained solution in EPEIRES is to use transparent bounded models and to map them (put, rotate and scale) on the real-life documents.

Mathieu: What are the recommendations to use this tool. **Epeires report 2006:** The groundtruth from real-life documents must have a least 5 pixel precision. Also, a groundtruth zone must contain a whole symbol. It is preferable to make bigger the ratio (groundtruth surface / symbol surface) than to miss a symbol part.

Mathieu: Also, how does work the checking process in EPEIRES, how you will control the groundtruth edit by someone.

Philippe: It is based on a verification step. An expert valid or invalid a groundtruth data edited by an operator. This means that there is no post-correction of this groundtruth by the expert, the operator has to re-edit it again.

Mathieu: Have you an idea on the needed time to realize such groundtruthing.

Philippe: No exactly, but I suppose that for a database composed of some thousands of drawing it will require several months.

Mathieu: I propose here an evaluation to discuss based on the EPEIRES model. One comment, there is two validations for each symbol to groundtruth (2 T_v and 4 T_v).

$$T = S \times \left(\boldsymbol{\alpha}\left(T_g + \mathbf{2}T_v\right) + \boldsymbol{\beta}\left(T_g + T_c + \mathbf{4}T_v\right)\right)$$

T:	global time
S:	number of symbol to groundtruth
α :	one shoot rate (α+β=1)
β :	error rate (α+β=1)
T_g:	mean groundtruthing time
T_v:	mean validation time
T_c:	mean correction time

T_g:	30 s
T_v:	5 s
T_c:	10 s
α:	0.85
β:	0.15
S:	100 000
y =	60s*60m*35h*52w = 6 552 000s
T =	4 300 000 s / ie. 0.656 y

Philippe: I agree with this formulae but may be we could add a constant time K related to general groundtruthing process (to open/close the software, to download the images from the server, to upload the results, to synchronize the data, ...).

Mathieu: I agree, but we can also consider this constant K as a part of T_g, T_v and T_c.

Mathieu: In the future may be some idea to speed up the process could to use a user driven segmentation algorithm (i.e. one click on the symbol will put automatically the box).

Philippe: Yes, what we have to do also in the future is to use spotting systems to help in the groundtruthing process.

Mathieu: Have you thought also to work on the GUI ergonomics, it could reduce a lot the groundtruthing time *Tg*.

Philippe: It is the second main perspective with the works on the user driven groundtruthing (i.e. using a spotting method).

2.2 Synthetic approach

2.2.1 Floorplans

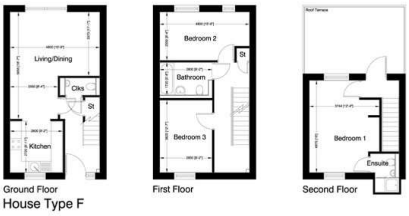
Jean-Yves: Using a similar background for a given database will give the possibility to the people to learn the background in order to separate the symbol layer.

Mathieu: The main purpose of the performance evaluation is not to find the best method but to find the methods' weaknesses. Using previous knowledge concerning the type of background has no interest, we can't take into account in this work the cheaters.

Ernest: That we can do is to combine different backgrounds in each test (with an unknown number of background for the participant). Next if the tests are done during a day-contest it will be very difficult for the participants to do such cheats.

Rashid and Jean-Yves: Having a background composed of thick line makes easier the localization of symbols. Why not using the same thickness for all the parts ?

Mathieu: The first floorplans we have generated has been done by using real documents found on the web. On these documents most of them are drawn with thick walls (see below).



Dimosthenis: The textures are they relevant of the architectural domain?

Marçal: Many floorplans use textures to draw the walls or the floors, it could be a good idea to define some textured backgrounds.

Mathieu, Ernest: Not all the methods work in the same way, the background type could have a great impact on the spotting/recognition results. Some methods could work from skeleton and in this case thick lines could introduce lot of noise. Other methods work from loops or connected components and textures will introduce false ROI (Region Of Interest). At last, methods working with contours will meet problem with empty wall (only regular width on the document) or texture (line of weak width). In order to test all these aspects we propose to generate floorplans using different kinds of background (filled, empty and textured).

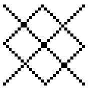
Hervé: The evaluation system has to evaluate the symbol recognition & spotting, not a complete analysis system. It is quite natural to implement a filtering technique in a system to segment the thick and thin lines, or any pre-processing tools dedicated to other tasks. However we are focusing here only on the symbol recognition & spotting. Using filled, empty and textured floorplans is an excellent idea.

Mathieu: One problem is to know how to use these different background types. Must we use three versions for a same background (filled 'i.e. solid', empty, textured) or different filled, empty and textured backgrounds.

Dimosthenis: That we can do is to compute systems' results for each ground. Like this we will be able to compare these results by level of ground difficulty (from the simple to the hard ones). This will not introduce a domain dependant criterion in the evaluation and will allow to analyse the impact of ground type on the automatic processing.

Ernest: The evaluation framework must be general enough to be use with different domains. I think we have to be careful to not be too domain dependant in our tests. In real-life floorplans you will never find a same document using different background styles (i.e. solid, textured or empty).

Mathieu: Concerning the empty background the key idea is to have a same width for the symbol and background layer (so it is 'easy' to do). Concerning the texture every body are they ok with the one proposed here, must we use other ones ?



Marçal: There is lot of texture in the floorplans, but it is very difficult to find common classes of texture. There is lot of different styles, this depends of the feeling of the architect when he draws. Also, we have textured symbols and floors on the drawings (not only the walls).

Hervé: Concerning the textures, we have to think about a pattern using either the same thickness or not of symbols (I agree with the Jean-Yves and Rashid).

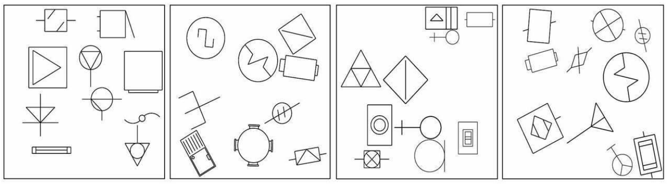
Hervé: I suggest also to add a “dimensioning layer” which may be easy to add within the Mathieu's groundtruthing application: defining arrows, and, as a constraint, a text sliding on it. In the same way, a label can be written, without overlapping with an existing symbol for each room. To sum up, I suggest to define a new kind of symbol corresponding to text. We can also think about dimensioning arrows overlapping both the background and the symbols (but the text).

Ernest, Mathieu: We can propose a large set of floorplans tacking into account texture (wall, symbol, floor) and meta elements (text and dimensioning). However, if we want to start soon the evaluation campaign we have to limit the elements to include in order to generate first databases as soon as possible. We propose to use at this time only texture in the walls (the classic ones 'grid and hash'). We will see next to include other elements.

2.2.2 Bags of symbols

Rashid, Jean-Yves: These bags of symbol are not realistic and the localization is too easy.

Mathieu: The key idea with these databases is to propose an intermediate level between the past contest's documents (isolated symbols) and whole drawings. It will allow people working on segmented symbol recognition to shift to symbol localization. I assume that not all the people will be interested to work on (especially people already working on the spotting), but this will help some others to adapt there past methods. Next these databases are also more adapted to test the scalability of the methods. Indeed, building documents with large model sets '25,50,100,150' is possible with the bags (not with the drawings, floorplans or diagrams in which the symbol models are imitated to a single domain). Also, the bags allow to test the invariance to scaling and rotation of the methods. The next figure presents some example using or not rotation and scaling. In order to make harder the localization we plan to use the degradation methods of the past Contest 2003 and 2005.



(a) (b) (c) (d)

Jean-Yves: Scalability will be tested by changing the number of different models in each test?

Mathieu: Yes.

Hervé: We have to keep in mind that some symbols can be disconnected, making a component labelling algorithm not the solution to spot symbols. Degradation methods can also make symbols disconnected, or connect symbols together.

Mathieu: Yes, especially because we have generated these documents using a particular positioning algorithm (the insert algorithm, see these slides). The result is the symbols are very closed (some ones are spaced of only some pixels) which will make many connections between the symbols with distortion methods introducing binary dilatation (as we can do with Kanugo). We have now to define the settings to noise the symbols (the ones of 2003 and 2005, new ones, mixed the noise types, how many level 3,6,9) ?

Jean-Yves: I agree. But it means that the number of disconnected and/or degraded symbols should be significant in such dataset in comparison to other ones. Why not using other kinds of symbols to change a little from the used and used architectural and electrical symbols? I think it is not difficult to use different symbols in the bags of symbol? Why not to introduce more symbols with filled shapes? disconnected parts? dotted lines? and so on? (I have sent a CD with a set of mechanical symbols to Philippe for the EPEIRES project).

Mathieu: At this time we have tried to match to the current and past organisations of the Contest. No new symbol models have been added in the "official" databases so we have kept the existing ones. However we could introduce new symbol models if every body agree. I propose to not change the existing model sets of the past editions (25,50,100,150) but to add a new one (upper to 150) including the new symbols of Jean-Yves. Another idea could be to do bags composed of line based symbols and bags composed of region based symbols. Please tell me?

Hervé: New symbols have to be used, no matter we are already able to deal with them. I think performance evaluation both has to highlight what we are already able to spot and challenges. Till now, it is right that filled shapes and dotted lines were missing. During the EPEIRES meeting at Rouen, Jean-Yves has also spoken of what I call "regular symbol", I mean symbols that are not rigid (i.e. we cannot fully define the instance of a symbol from 3 non collinear labelled points). Nevertheless, introducing such symbols within the actual framework will be a difficult task (should be a challenge point within the conclusion of this study).

Hervé: Bags of symbols are the solution we can keep to evaluate scalability of the methods without thinking about a domain application which involves many symbols. Moreover, while we are able to evaluate true detections, we can also evaluate false detections which may occur from various parts of various closed symbols (very closed / connected when adding noise).

Mathieu: At this time I suppose we agree on the key idea that the bags of symbol are a good way to test the scalability (with a localisation objective) of the systems in comparison to the floorplans. They are also a good deal to test the localization abilities of a system in different way of floorplans.

Ernest, Philippe: The main objective is to test the localization. The bags are intermediate databases allowing mainly to test the scalability of methods. Also, the purpose of the evaluation is to be real as possible. The evaluation framework (and then the used benchmark databases) must simulate real contexts, and this cannot be done with the bags. So to conclude the bags are interesting databases but not the priority ones. We can keep them in mind but the priority is to test the localization on the whole drawings (floorplans, diagrams, etc.).

2.2.3 Degradation models

EPEIRES report 2006: Concerning the degradations, the 2005 GREC contest has shown the limit of a method such Kanugo. Some participants have previously learned the parameters of distortion methods and next applied a reverse transform (i.e. full cleaning).

Mathieu: Any kind of mathematical model will produce a deterministic noise, more or less easy to remove by a system when the method type is previously known.

Ernest, Philippe: What we propose to do for the future contests is "blind tests". We will mix the noise levels (i.e. different parameters) within a same test. Next we will combine also different methods together (Gaussian, Kanungo, etc.). Like this it will be very difficult for the participants to use previous knowledge concerning the distortion methods and their associated parameters.

Jean-Yves: Some distortions of the past contests were too strong and not realistic. Testing the recognition systems on such degradations has no interest. We have to keep in mind than in the real-life documents we will never meet this kind of distortion.

Mathieu: Does exist some method allowing to produce more realistic noise than Kanungo?

Ernest: In fact the purpose of Kanungo is to do a realistic noise, it is becoming a kind of standard with the years.

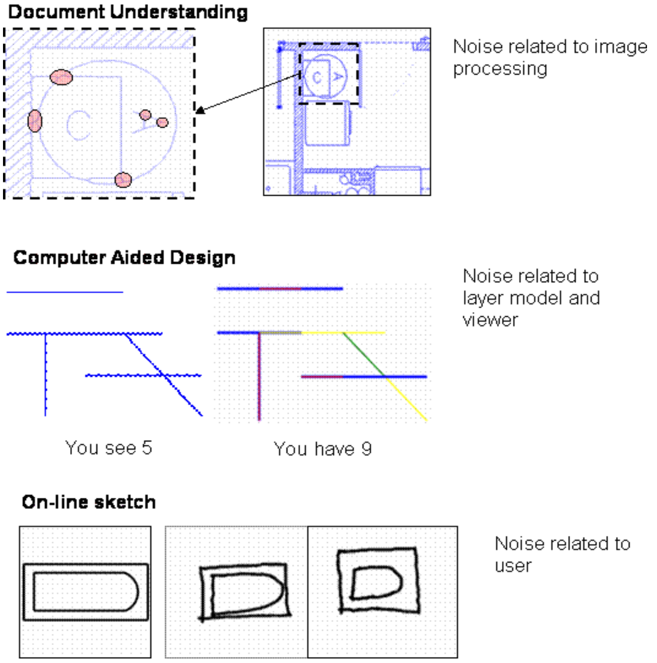
Ernest: Concerning the noise level I think we have to produce two kinds of dataset. First to produce large document sets noised in a realistic way (i.e. so with a reasonable level of noise) because it corresponds to the true-life. Next to produce small document sets noised in a unrealistic way (i.e. with a high level of noise) because we have to test also the limits of methods.

Jean-Yves: Yes, I agree with this proposal. We have to take care to propose realistic distortions but also unrealistic ones to test the limits.

Mathieu: Do you think that the realistic constraints will be still compatible with the combination of degradation methods you propose (i.e. blind tests).

Ernest: Yes, it is.

Marçal, Mathieu: Concerning the vectorial distortion we believe than three kinds of noise could be meet in the real-life:



Marçal, Mathieu: In regard to the purpose of the evaluation (spotting/recognition on printed documents) we believe that only the first one has to be used. But only in the case we propose tests at a CAD level (i.e. on vector data). Also, in this case it is not necessary to develop some new methods. Indeed it is possible to start from groundtruthed images and to apply a commercial vectorization tool: the images will be distorted and the groundtruth will stay still valid.

2.2.4 Sizes of images

Rashid, Jean-Yves: Having symbols of 2-3 pixel width is too weak and not realistic in the synthetic documents, what about the sizes of images and symbols.

Mathieu: Following different discussions with Ernest and Philippe we have decided to take a bounding box of 256*256 per symbol (the half size of the 2003 and 2005 Contest editions) which corresponds to a width of 4.5 pixel per line. We cannot do more because it will produce too bigger images (not all the methods have an extraction process fast enough).

Jean-Yves: Can you explain why the sizes of images must be limited, I don't understand? only to limit the processing times?

Philippe, Mathieu: It is important to limit the size in the case of day-contests. Having 'small' images allows to reduce the processing times. With a minimum symbol size of 256*256 we have fixed an upper limit of about 10024*1024 pixels for the bags and 4000*4000 pixels for the floorplans (by considering only the backgrounds composed of a weak number of rooms, from 4 to 8). Another key point is than not all the systems are stable enough to process large images. Some ones use too complex algorithms to extract the features or for the matching process. Some others could have technical problems that will produce 'overflow memory' errors when they process large images. We must keep in mind that finding participants for a Contest is a harder task (see [4]). It is necessary to allow every people to participate, not to discourage them.

Rashid, Jean-Yves: Why have you choose to change the size of symbols on the floorplans?

Mathieu: On whole drawings it is obviously necessary to use symbol having different sizes. We cannot keep the same size for a table, a tub or a chair. It is not realistic because not well proportioned. So I have taken a size of 256*256 for the lower ones and the bigger ones depend of the application (at this time 614*614 for the floorplans see [5]).

2.3 "A priori" knowledge

2.3.1 Introduction

Jean-Yves: In the different localization tests of the contest, what will be the data provided to the candidates for each test? only a set of complete images without any information about the used symbols? model images of the different symbols will be provided for each test (learning step is possible)?

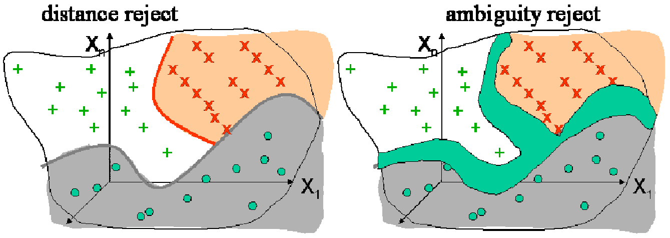
Mathieu: These previous data should depend of the application to evaluate: recognition or spotting.

- **Recognition:** In this case the symbol models are off course required by the systems in order to provide their labels. These models could be ideal symbols or real samples (in order to do a learning step).
- **Spotting:** Concerning spotting most of the systems work with Query by Example (QBE). So what is needed are images of segmented symbol in order to use them as queries.

I propose to discuss these two points in the next subsections.

2.3.2 Learning data

EPEIRES report 2005: Concerning the recognition it is important to provide real samples for the methods working with learning, not only the ideal models. It will allow to compute reject rates of classifiers. However, we have to take care to include the full list of images (he training and test ones) in the groundtruth to compute these rejects.



Jean-Yves: I agree, but we must also know the sizes of learning databases used by the systems. In the field of Pattern Recognition and Machine Learning it is a very important issue. The size of the learning database has a great impact on the recognition results. Doing recognition, without having precise idea on the size of the learning database, has none sense. Until now, I don't understand why we haven't integrated this consideration in the past Contests. For example statistical methods could use more learning data than test ones before to be efficient (see Machine Learning Repository). In another hand, the structural ones will often unable to use more than the ideal objects (structural learning is a complicated process).

Philippe: I agree with this comment, the participants have to mention explicitly if they full use or in full or in part the training databases.

Jean-Yves: Also, we could extend the evaluation tests by tacking into account several training/test ratios (e.g. 10% training 90% test, 20% training 80% test, etc.).

Philippe: I not agree. We have to keep in mind that we want to evaluate here is the localization, not the classification (that is the purpose of the Contest's part on the segmented symbols). Moreover, more we will increase the number of requirement for the localization evaluation, less we will find participants. We have to consider that not all the systems work with a learning step, most of the ones working at a spotting or segmentation level rely on a structural approach. These systems don't use, most of the time, a learning step. So we have to take care to not do an evaluation that will be system dependent. However, I'm not totality opposed to this idea. I just argue here that we have to do it only if it becomes a priority task in the work concerning the localization evaluation.

Jean-Yves: Another comment concerns the way to do the learning. In the past Contests the learning databases were provided before the test data. I think we must link these data (in a single package) for the future. The methods should do their learning at the beginning of the Contest (not previously).

Mathieu: I agree, at the end it was difficult to know which model dataset we had to use. However, the case was different because only ideal models were used in the past Contests. So, to split the model and test data allowed like this to limit the amount of data to download (e.g. when a model set is used for 9 different tests, it is then reasonable to download it a single time). But, in the case of real samples the case is different: the learning dataset has to change for each test. So to link these data together is a logical thing.

Mathieu: I suppose that everybody agree with the idea to use real samples. But a question remains how to generate them? ideal crops form the groundtruthed images? using images generated randomly like in the previous contest edition? etc.

Jean-Yves: The second idea could be an easy solution to do it.

Hervé: I not agree, the localization Contest is not the recognition one. We have to train on contextual data (i.e. extracted from whole documents), not from images of segmented symbol generated randomly.

Jean-Yves: Yes, but in this case we must use different drawings to provide the learning and test data. It will be not reasonable if the learning data was included in the test one.

Hervé: I agree, but we must take care to keep coherency between the both. This means that the training data must relate to the test ones. However you will learn nothing.

Mathieu: If we extract the images from whole drawings, is the number of sample must be same for all the classes?

Hervé: No, for the training the number of sample per class depends of what we have in used data. So it is proportional to what you have found.

Hervé: Another comment concerns the learning of contextual information. The systems have to learn, off course information concerning the models (i.e. connected symbols). However, they can also learn other contextual parameters (graphical charts, textures, width of walls, etc.). The best way to constitute the learning databases is then to do give to the participants the whole images with the associated groundtruth (not only the cropped images). Like this the systems will be able to learn as soon as symbol samples and contextual information.

Mathieu: Yes, but in the case of synthetic documents, do I have to give samples for each background?

Hervé: When I say "to learn context" I think to learn the graphical charts of documents. Off course, this doesn't mean specific stuffs related to the performance evaluation framework (like the background types, the noise method and its parameters, etc...).

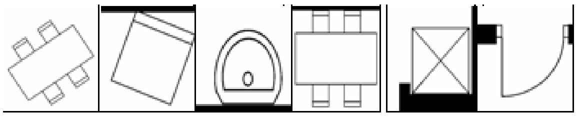
Mathieu: So, one idea could be to take at random whole images from the document database.

Hervé: I agree with that.

2.3.3 Query by Example (QBE)

2.3.3.1 QBE generation

Mathieu: Concerning the QBE the first thing is to decide how we will generate them. That we can do is to start from the images and their corresponding groundtruth (i.e. the bounding boxes or others wrapper objects) and next to apply an algorithm to crop the symbols in a random way. This algorithm could work in two steps: first shifting (i.e. to translate the bounding boxes) next sizing (i.e. to change the bounding boxes' sizes).



Jean-Yves: Yes, but from my point of view we have to include other kinds of QBEs, I have listed three ones.

- Using just the ideal symbols.
- Using symbols in context.
- Using drawing crops (i.e. a random part of a drawing that will include more than a symbol) in order to test the fall-out (see this Wikipedia page) of systems.

Mathieu: In order to understand well the comments of Jean-Yves I remind here what the differences between the precision, the recall and the fall-out are.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

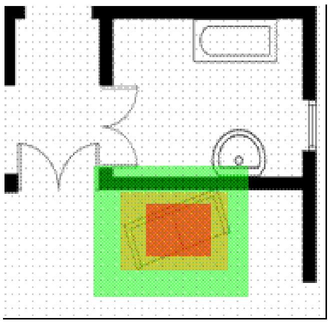
$$\text{fall-out} = \frac{|\{\text{non-relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{non-relevant documents}\}|}$$

Hervé: Concerning the fall-out I have a set of comments. First we must take care to not do confusion between the symbol spotting and the drawing retrieval [28]. Using large crops will shift our QBEs to document retrieval ones (i.e. to retrieve drawings having similar layouts). Next, at this time I don't see how we will evaluate the fall-out. Indeed, we have no groundtruth concerning the background (i.e. the non-relevant information), only groundtruth concerning the symbol areas.

Mathieu: Yes, you're right. But if you have the groundtruth *S* which describes the symbol areas you can compute /*S* which will describe the no symbol areas.

Hervé: I not agree, you have no information concerning the document background in your groundtruth (for you it is only a "black hole"). We have not to do queries about something whose we don't know the groundtruth.

Marçal: For the symbols in context you have to be realistic with your automatic cropping method. Most of the users do unstinting crops. That means they still try to have the full symbol area, including more or less background in the QBE. The cutting cases of symbol in the QBE are scare.



probabilities
high, low, null

Hervé: I'm ok with Marçal, I suggest here to learn previous parameters from users' samples before to generate the QBEs. We mustn't produce too unrealistic QBEs. Also, we can use other geometrical shapes to process the QBEs like the ellipsis, the polygons, etc.

Mathieu: Concerning the parameters there is previous work of Ernest we can use [29]. We could develop something adapted for the QBE case, and supporting different geometrical shapes.

Jean-Yves: Yes, but that you can do also is to combine the QBE difficulties: only easier, only hardest, easier and hardest.

Hervé: I agree with Jean-Yves, we can build several qualities of QBE and to combine them. However, due to the contest organization may be a part of these tests could be made optional.

Philippe: I agree, the QBE qualities could have a great impact on the evaluation results, we should think about that. However, I recommend here to build realistic QBEs with weak distortions. The performance evaluation must simulate realistic contexts, we have to adapt on the real use cases.

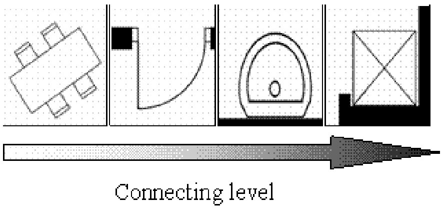
Mathieu: Yes, but in the other side one key objective of the performance evaluation is to test the systems' limits. However I agree we have to generate QBEs which keep meaning: if a crop misses totally the target symbol so the spotting system won't be able to query it. In that sense, our "threshold" is different from those of the binary distortion where symbol shadows still remain in the case of strong noise.

Jean-Yves: Also, I repeat here the same comment than for the previous learning section. You must use different drawings to provide the QBEs than the test ones. It will be not reasonable that the QBEs were included in the test documents.

Ernest: Ok, as for the binary distortion methods we have to define a method to generate the QBEs. It "should not" so complicated to do.

2.3.3.2 QBE number

Mathieu: The other problem concerns the number of QBE we have to provide for each test. Until now I think we agree that more we will propose QBE, more we will refine the evaluation. In all the cases, we have to produce more than one QBE per class. Indeed, a "hardest QBE" could weight too much in the spotting results of a method. To avoid such case it is necessary to take use more than one QBE per class.



Ernest: I agree, we should generate more than one QBE per class. The question is to know this number.

Jean-Yves: From my point of view it is a statistical question. We must see from how many QBE we will be able to obtain good statistical analysis.

Hervé: We have to take care, to use a large number of QBE will raises a complexity problem. Indeed, the full evaluation will be achieved by computing a rank for each QBE. So, considering a database composed of *n* symbols the total number of comparison to do will be very high as shown in the next figure (I consider here the bijective distance case). Using a random process to select a subset of QBE from the database will limit this complexity.

	1	2	3	4
1	0	a	b	c
2	a	0	d	e
3	b	d	0	f
4	c	e	f	0

$$N=\sum_{i=n-1}^{1} i=\frac{1+(n-1)}{2}(n-1)$$

n=100 N=4950

n=250 N=31125

Marçal: It is not necessary to do all the comparisons for each QBE. In my system [18] I use a Hash Table (HT) to index the database. The extracted features from the QBE allow to obtain directly the rank list from the HT.

Mathieu: The HT could be use in the case of the statistical approaches to structure your index but not for the structural ones like in the Hervé and Rashid's systems [19] [20]. So in this case the number of QBE raises a complexity problem.

Jean-Yves: I'm not sure that the number of comparison for the spotting is bigger than for the recognition. For the recognition you must take in consideration the number of model per class. In the spotting case, you have only one object to compare (the QBE).

Mathieu: Following a phone discussion with Jean-Yves I have compared the complexities of the recognition and the spotting (see below). If you agree, at the end the complexities are same if the number of QBE equals the number of model. However, we have concluded before that we must have (at least) more than 2 QBE per model. So in all the cases the spotting will be twice more complex than the recognition.

$$N=QBE\times\sum_{i=0}^{j}S_i \quad \text{(1) spotting}$$

$$N=M\times\sum_{i=0}^{j}S_i \quad \text{(2) recognition}$$

$$N=N\Rightarrow M=QBE \quad \text{(3) condition of equivalence}$$

N:	number of comparison
QBE:	number of query
j:	number of image
S_i:	number of symbol of the i image
M:	number of model

Jean-Yves: The above formula uses the hypothesis that there is only one example per class in the test database. I disagree with that. The number of model per class is an important feature that could be used by the systems to increase their performance in a “malicious” way. We have to make differences between M = number of classes, K = numbers of models in each class and X = M x K = size of the learning dataset.

Mathieu: Yes Jean-Yves you're right. I have forgotten this important feature. I have updated the formula.

$$N=QBE\times\sum_{i=0}^{j}S_i \quad \text{(1) spotting}$$

$$N=M\times K\times\sum_{i=0}^{j}S_i \quad \text{(2) recognition}$$

$$N=N\Rightarrow M=\frac{QBE}{K} \quad \text{(3) condition of equivalence}$$

N:	number of comparison
QBE:	number of query
j:	number of image
S_i:	number of symbol of the i image
M:	number of class
K:	number of model in each class

Mathieu: However, as we have already explained previously most of structural methods use K=1: they are unable to do learning. So, in the structural case the complexity stays the same.

Jean-Yves: As I have explained before, the QBE number should to be chosen using only statistical consideration. However, if we have to limit the size of data for the spotting I agree to limit the number of image and symbol/image, not the number of QBE.

Ernest: I agree too, if we have to do a choice we must reduce the size of the test database in order to keep a large number of QBE. However we have to take care to not be method dependent in our performance evaluation approach. From my point of view a spotting system must deal with the complexity in the real-life (a web user doesn't wait a long time a query answer, he will browse to a next link). So the complexity is integral to the spotting topic. If a system cannot deal with this constraint, then it is not a spotting system but a localization one. The performance evaluation must be generic and defined according to real problems, not according to the programming level of existing systems. So the systems have to adapt themselves to the performance evaluation framework (which simulate as possible the real life problems), and not the opposite. For that reason we must make complex the spotting evaluation in order to simulate real-life conditions.

Hervé: I agree with the comments of Ernest. First it is important to use a large number of QBE to reduce the impact of a worst QBE on the evaluation. Next I agree too that a spotting system have to deal with the complexity. At the end, I think that in a same way of Marçal a graph based spotting has to introduce heuristic to limit the space search. So, the complexity in the case of spotting is not a relevant problem to consider from the performance evaluation side.

Philippe: I agree also with Ernest, the spotting systems have to deal with the complexity aspects. Moreover, concerning the evaluation side I'm not sure that having large test dataset will improve the evaluation results. In the case of spotting, if the system works on few documents so it will be able to work on more documents. That will impact the results of spotting is the content variability of the test database, not its size.

Mathieu: I not agree with this comment. As in the recognition evaluation you must have a sufficient number of document if you want to have good evaluation precision.

Ernest: Concerning the spotting I think now we should test with different sizes of test database. I agree with Philippe that it should not change a lot the spotting results. However, it will allow to study the abilities of systems to deal with the complexity. Because this complexity is integral to the spotting, we have to test it using different database sizes.

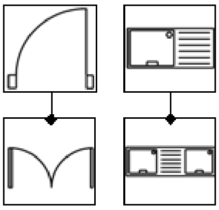
Marçal: In this case if we use only a QBE subset (from 1% to 99%) it has to be relevant of the database. e.g. this QBE subset must respect the distribution of symbol per class (if we have two doors for one table, so we must find this ratio in the QBE subset). The rotation gaps of symbol classes must be also respected (if a model is often rotated in the drawings, we have to provide a same proportion of rotated QBE for this model).

Jean-Yves: I not agree, because you have many pictures of Bill Gates on the web you perform all the days such query. The QBE mustn't be correlated to the test database.

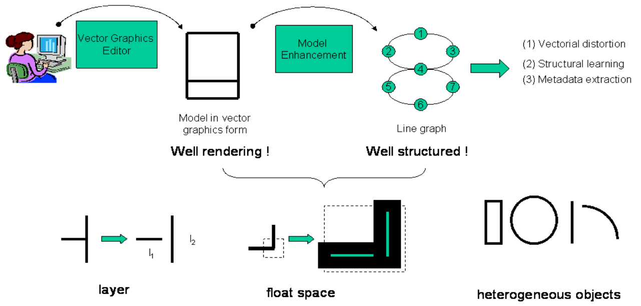
2.4 About scalability

2.4.1 Model scalability

Mathieu: At this time we have a problem concerning the groundtruthing from whole documents: the management of the model database. Indeed, when you define the groundtruth (as soon as in the real and the synthetic approaches) you have to browse in a fast way in your model database. When this database is composed of few models it is not a problem. However, when you have to deal with hundreds and hundreds of model, the spent browsing time could be high enough to make too harder the groundtruthing process. We have to define a system to browse in a fast way in the model databases using filtering on the graphical primitives, matching process of composition relations, similarity criteria between the models, etc.



Mathieu, Alicia: The key problem to do such process is the weak quality of actual vector graphics models. Indeed, they are edited using with vector graphics editors: so well rendered but not well formed. We have to develop a model enhancement step in order to correct the mistakes (e.g. to remove the overlapped objects due to the layer composition, to find the junctions between the lines and the closed objects "circles, squares, etc.", to split the lines, to connect the near extremities, etc.).

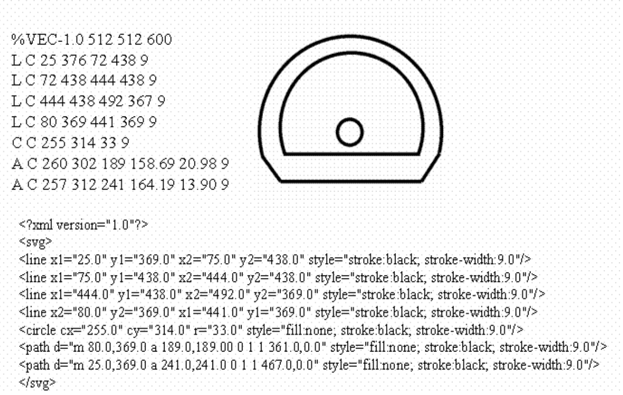


Philippe: From my point of view we're talking about two levels of process: how to make connect the graphical primitives and how to retrieve the models. In all the cases the first one is important to discuss for the methods' learning (e.g. build line graphs from the vector data).

Mathieu: This enhancement process is also needed for the computation of anchor point (i.e. lines' ends, junctions, etc.).

EPEIRES report 2006: These anchor points are requirements of industrial people, but it is certainly a very specific need that has no relation with the performance evaluation.

Mathieu: Another problem related to this process concern the VEC format. This format cannot preserve the connections between the points (especially because the arc objects are represented using starting and ending angle). With the SVG format we have no preservation problems.



```
%VEC-1.0 512 512 600
L C 25 376 72 438 9
L C 72 438 444 438 9
L C 444 438 492 367 9
L C 80 369 441 369 9
C C 255 314 33 9
A C 260 302 189 158.69 20.98 9
A C 257 312 241 164.19 13.90 9
```

```
<?xml version="1.0"?>
<svg>
<line x1="25.0" y1="369.0" x2="75.0" y2="438.0" style="stroke:black; stroke-width:9.0"/>
<line x1="75.0" y1="438.0" x2="444.0" y2="438.0" style="stroke:black; stroke-width:9.0"/>
<line x1="444.0" y1="438.0" x2="492.0" y2="369.0" style="stroke:black; stroke-width:9.0"/>
<line x2="80.0" y2="369.0" x1="441.0" y1="369.0" style="stroke:black; stroke-width:9.0"/>
<circle cx="255.0" cy="314.0" r="33.0" style="fill:none; stroke:black; stroke-width:9.0"/>
<path d="m 80.0,369.0 a 189.0,189.00 0 1 1 361.0,0.0" style="fill:none; stroke:black; stroke-width:9.0"/>
<path d="m 25.0,369.0 a 241.0,241.0 0 1 1 467.0,0.0" style="fill:none; stroke:black; stroke-width:9.0"/>
</svg>
```

EPEIRES report 2005, Philippe: Concerning the vector graphics format the Contest participants like the VEC format because it is easy to parse. However, its representation properties are limited. We have to define some extensions concerning the filled shapes, the circles, the polylines, etc. Some proposals have been done in the past, we have to group them together and to do formal proposal of VEC 2.0.

Mathieu: I think we must be careful by extending the VEC format. It is not necessary to develop a new SVG. The key property of this format is to be simple as possible. So, if we want to extend it too much, then we will make it too complex. In this case it seems better to use directly SVG.

Philippe: We agree, the VEC format must stay simple as possible. However it could be updated for improvement (like to correct the "arc problem"). My felling is we have to keep the VEC format because the SVG one is to hard to manage for the users (especially the path tags). But if there is a data preservation problem we have to change it.

Mathieu: I agree the SVG is hard to parse. However that we can do is to use the SVG from the evaluation side (the management of the model database) and to provide VEC data to the participants. Like this, we will take benefit of the available SVG editors. Moreover, we have already some SVG/VEC converter.

Philippe: I agree, only the evaluation framework has to deal with the SVG. It could be a solution.

Philippe: However, concerning the general browsing process we're talking about, a problem still remains. Not all the models are in a vector graphics format. Especially the ones extracted from the real documents. So, at then end if we develop a vector graphics based browsing we won't be able to browse all the models. I propose in that case to investigate the use of raster based signatures. Indeed, you can easily convert vector data into raster. We can study some well known methods like the Fourrier Transform or the R-Signature.

Mathieu: Yes you right, but with a raster based signature you will be able to browse only at a shape similarity level. You won't deal with the primitive filtering or the composition relations. To do that, you have to handle vector graphics models. Ways to solve this problem are to apply a vectorization step or to re-edit vector graphics data from the bitmaps.

Hervé: In all the cases, to use a raster or a vector graphics based signature, you will still have the problem query initialization. How to start the browsing when you have to use QBE to describe a wished model? Another possibility to retrieve the models is to use hand-sketch queries [27].

Ernest, Philippe: Not all the people could have plot tables. Another simple solution is to add metadata in a hand user way from models. There are no so much models (some hundreds), to edit these metadata won't take to much time. We can already do that in the EPEIRES platform, each model is dealt by couples "property + value" (you can propose several couples for a same model). You will find more details about the used process in this publication [24].

Mathieu: The metadata could be a solution when the groundtruther (i.e. people doing the groundtruth) is skilled in the domain. Indeed, he must recognize the symbol (to know its name "i.e. the metadata") in order to retrieve it using keyword(s) in the model database. In the case of complex domain (with a high number of model) it becomes harder to do it. The groundtruther could then waste lot of time to explore the database by testing several keywords. The use of signature to retrieve the models seems to be the alone solution. However I agree with the Hérvé's comment, the problem is to initiate the query? This initialization problem makes this approach impracticable without using an hand-sketch interface. I suggest here may be another way to explore in order to overcome this hand-sketch solution, the use of graphics taxonomy [31]. The user should find its model by browsing in a shape tree or graph. The low levels should represent the basic primitives (arcs, straight lines, etc.) and the high ones the complex shapes (square, cross, diamond, etc.).

2.4.2 Domain scalability

Jean-Yves: At this time the EPEIRES platform has been dedicated only to the architectural drawings. It is a pity to limit the evaluation on a single application domain (the floorplans). Electrical diagrams could be also interesting for the evaluation.

Mathieu: The third edition of the Contest is a kickoff concerning the evaluation from whole documents. This constitutes an important gap for the systems and to limit the Contest to a single domain seems to be fair. We have chosen the architectural drawings in recognition to their interesting properties concerning the connectivity and the orientation of symbols. We can also consider the electrical diagrams, but it will be done in a second step.

Ernest: Off course for a full evaluation we have to include other document domains. But we have to start from a point first, architectural drawings is a good starting one. We will see following the first evaluation campaign to produce other domain databases.

Hervé: I agree, we have to finish the evaluation framework first (i.e. document, groundtruth and performance characterization). The evaluation of other domains could wait.

Mathieu: One important point to highlight is that the participants consume the tests. I mean, once the tests used they expect other tests in order to proof their works, without any return on the used tests. At the end we know now it will be very difficult in the future "to refresh" the real-life tests due to the time delay to constitute them. The "production" will not follow the "consumption". Concerning the synthetic approach the problem is to evaluate the quality of the produced databases. The only criterion we have is to compare the systems' results on the real and on the synthetic databases. So, in the both cases it will be necessary in the future to combine the synthetic and the real-life approaches. The first one allows to produce databases with more variability concerning the content. The second one allows more flexibility in the production, and to deal with large amounts of data. So we have to take keep coherency within their combination (i.e. if the real-life databases are architectural then synthetic ones must be same).

Marçal: Also, one advantage of architectural drawings is their rich graphical symbols. The Electrical diagrams have less expressive symbols (i.e. I mean that they are composed of few lines, loops, etc.). In that sense, it seems more easy to do the localization/spotting from architectural drawings. It will encourage the participants to do the Contest.

Jean-Yves: I agree, the architectural symbols are more simple to discriminate on the images. But the electrical symbols present a high level of variability in regard to the architectural ones. They contain filled primitives and belong to a large number of model.

Jean-Pierre: I agree, the electrical diagrams present better properties to test the scalability of methods than the architectural drawings. This is a large number of models (several hundreds).

Mathieu: Yes, but in regard to the model scalability topic you agree that we won't be able to deal with such number of model in the electrical diagrams.

Philippe: We have to take in mind that our purpose here is to test the localization, not the classification as in the past Contests. So scalability in the case of whole drawings is not a priority objective. However, I think this scalability will be not a problem with the PIVERT platform. We have already tools to browse in the model databases (tags property + value). The remained problem is the groundtruthing time. At this time the EPEIRES platform has been used only for the groundtruthing of floorplans. It seems difficult to include other document domains in the next months (we have already overflowed the deadlines). So it is better to build only synthetic floorplans in order to compare the systems' results as soon as on the real and on the synthetic data. For the future we plan obviously to include other document domains, and obviously electrical drawings. This will change the localization context, so it is a "next priority objective"**.**

Mathieu: Ok, so in this case this scalability difficulty remains only from the synthetic side. We have to integrate similar browsing method in the 3gT system in order to deal with it.

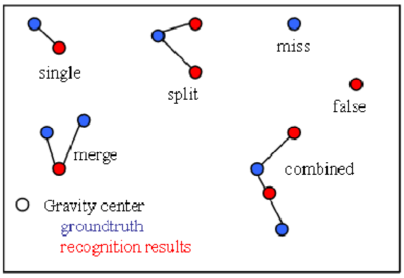
Jean-Yves: Ok, but in all the cases having synthetic electrical diagrams with 20 models is better than having no electrical diagrams at all (and only architectural drawings).

Mathieu: I can produce synthetic electrical drawings, but every people must agree that we will have only synthetic documents (not real ones) with a low number of model.

3. Performance Characterization

3.1 Introduction

Mathieu: in the case of symbol localization the characterization becomes harder because the matching between the ground-truth and the results is done between object sets (i.e. gravity centers, bounding boxes, regions, etc.). These object sets can be of different size, and large gaps can also appear concerning their localizations. As defined by [6] five matching cases can be done, we show them in the next Figure: single, split, merge, missed and false. Moreover, combinations can also be done within the split and merge cases.



Mathieu, Ernest, Marçal: These cases are the basis of the characterization of the localization. Their detection makes possible the computation of the recognition rates and the ROC curves (i.e. precision vs. recall). The remained problem is now to detect these cases from the groundtruth and the systems' results. Several approaches could be used according to the handled data types (point, bounding box, polygon, etc.).

Jean-Yves: Question, who will be able to realize the algorithms for the comparison (i.e. matching) of the groundtruth with the provided results? It could be very difficult according to the previous choices.

Mathieu: Yes you're right, but we have to try it. I have already realize a "near" programming task in the past [30], may be I could try again.

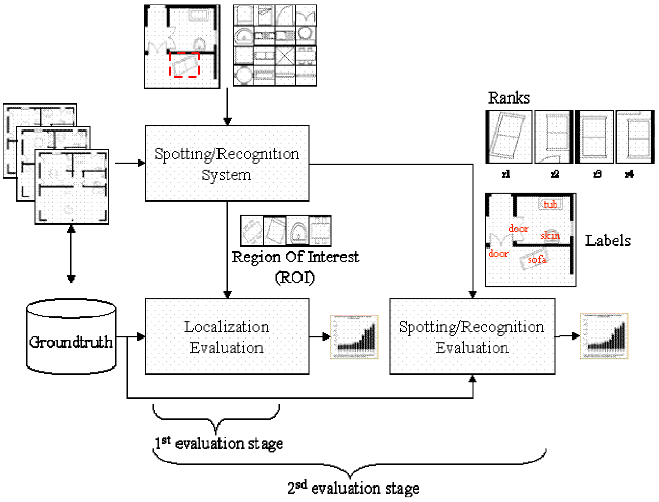
Marçal: I'm working on such task at this time, I have a paper in progress.

Mathieu: Concerning this problem I have a first comment: the localization cases (split, merge, single ...) will change the ranking and the recognition results. I summarize the problem on the next table.

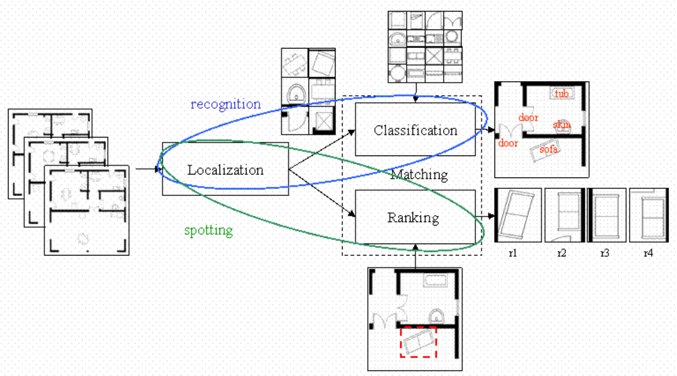
GT	Res	Pre	Rec
1	1	=	=
1	n	-	=
n	1	=	-
1	0	=	-
0	1	-	=

= (no variation) - reduce the precision or the recall

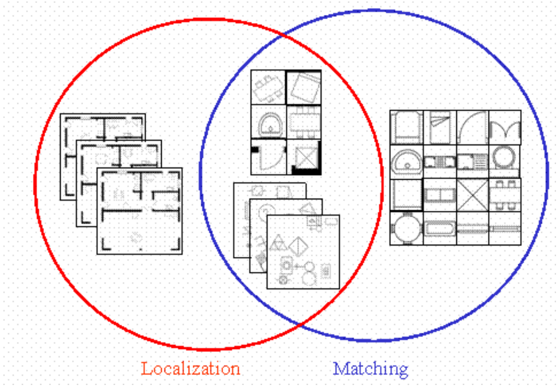
Mathieu: The question is to know now if we have to evaluate in one or two steps. Indeed, the localization results could change a lot the matching ones. May be it could be good thing to evaluate these two levels in a separate way?



Mathieu: Then, if we consider all the possibilities we have different evaluation scenarios: localization, classification, ranking, recognition and spotting.



Mathieu: Obviously, we will have to use different benchmark databases. Indeed, a given database could have good properties for a given task, but not for another. At this time, I think we can consider two main database properties: localization and matching. The whole drawings (like the floorplans and the diagrams) have to be use to evaluate the localization. The segmented symbol databases present better properties to evaluate the matching (classification, ranking). Indeed, we have no localization step, the methods can be evaluated only on the matching results (testing the scalability, separability, etc.). The bags of symbol allow to test the methods on the localization and scalability criteria, so they correspond to "hybrid" databases. At last, the cropped symbol images are similar to the segmented symbols but with parts of background.



Hervé: In any case, if we evaluate the localization in an independent way from the matching, we have to use the same prior knowledge. I mean, to use the QBE for the spotting and the model databases for the recognition. At the end we will test the localization results, but using a spotting/recognition context.

Marçal: I not agree with the idea to evaluate the systems in two steps. You have to define an evaluation protocol in which you will take into account as soon as the localization and the matching results in one shoot.

Philippe: I agree, we have to keep in mind the final goal of our evaluation: the recognition and the spotting. The localization is not a priority objective, it is just an intermediate level (e.g. if you evaluate the localization why not to consider other tasks like the vectorization or the image pre-processing). At this sate of our evaluation framework I think this will be an error to consider such "extra" criterion. Moreover, I think we have to study carefully the relevance of this criterion before. At this time I don't see any positive argument to valid it. If we use it, I propose to keep it for the evaluation people (not to make it public). We will be able like this to study its relevance.

Mathieu: Yes, but some systems are interested only in the segmentation [20] [21], then how to evaluate them?

Ernest: I agree, the spotting systems are often black boxes. So in this case the one step evaluation seems to be the best choice. However, some recognition systems work in two steps (localization and then recognition). So, for these systems we have to propose a two steps evaluation. I propose here to do the following evaluation:

- localization only
- localization + classification (i.e. recognition)
- localization + ranking (i.e. spotting)

Jean-Yves: I'm ok for to evaluate the localization, spotting and recognition independently (3 types of tests).

Hervé: Same.

Jean-Yves: I have some additional questions/suggestions about this point:

- For the localization we have to evaluate differences between positions and sizes of Regions Of Interest (ROI) in the groundtruth and in the systems' results. So question, what is a ROI: a point ? a bounding box ? etc.
- For the spotting we have to evaluate the pertinence of the ranking returned by a system. So question, what is a ranking: an ordered list of image name ? an ordered list of image names + degree of pertinence ? an ordered list of regions of interest with image name ? an ordered list of ROI with image name + degree of pertinence ? etc.
- For the recognition we have to evaluate the pertinence of the whole document interpretation. So question, what is an interpretation: a CAD file ? a SVG file ? a list of symbols ? a list of symbols with confidence rate ? I think we will choose the last item ! So question, what is a symbol ? a point ? a bounding box ? a raster image ? a SVG description ? a CAD description ? etc.

Finally, about this point, my opinion is that it will be easier to generate different groundtruth for each type of evaluation (localization, spotting and recognition).

Mathieu: I agree, we have to provide different groundtruth to evaluate these different applications.

3.2 Localization characterization

3.2.1 Groundtruth is points.

Philippe: Based on the EPEIRES discussions the groundtruth of the localization corresponds to points.

Mathieu: How will we decide of a possible overlapping between a result and the groundtruth in this case? I mean when you won't have equalities between two points.

Ernest: We have to use automatic thresholding methods to match the points together (based on a clustering of distances between the points). Another possibility is to use a sliding threshold, and next to display the different localization results according to the different sliding values.

Philippe: One problem concerns the distance to use between the points: Euclidean, min max, etc.

Philippe: Another problem is the impact of the sizes of symbols on the localization errors (ig. these errors will change a lot between symbols of 50*50 px and others of 200*200 px).

EPEIRES report 2006: Using only a point in the groundtruth makes impossible the evaluation of the scaling results provided by the recognition methods.

Mathieu: Moreover, using thresholding techniques with points will make impossible the identification of missed and false cases (it could associate very far objects that will be confounded as single or split cases).

3.2.2 Groundtruth is regions

3.2.2.1 Groundtruth is wrappers

Tony: The focus here at the moment to be on matching the ground truth and results. I think that's fine, but I wondered if there was some way to put error bounds or a probability distribution around the two types of data elements, rather than just thinking of it as point-point matching. Could we make some measures of e.g. the size of the inked regions around ground truth and reported results in the original drawings, where original drawings exist?

Marçal: I agree, to solve the problems of point to point matching we have to use region data in the groundtruth.

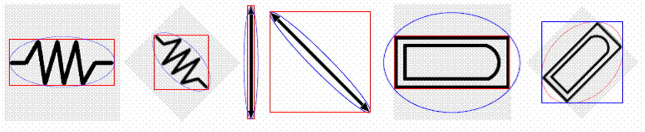
Jean-Yves: Why not use bounding boxes.

Mathieu: The thresholding problem is solved in part if we use bounding boxes in the groundtruth. Indeed, we can use the outlines of bounding boxes as natural thresholds. At this time this solution is possible in the EPEIRES system (bounding boxes are defined in the groundtruth).

EPEIRES report 2006: An including test of point results could be define from the bounding boxes

Philippe: There is only points in the EPEIRES groundtruth (upper corner and lower corner), but we can reconstruct the bounding boxes from these points.

EPEIRES report 2005: The bounding boxes are, may be, not the best way to outline the symbols. Some other geometrical shapes (like the ellipsis) could have better properties.

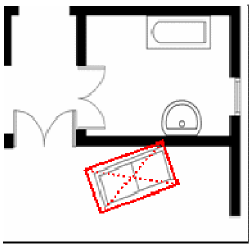


Mathieu: I propose to talk here about wrapper objects?

EPEIRES report 2005: A problem remains in all the cases, according to the orientation, the bounding boxes could match in better way than the ellipses (see the tub below). May be we will have to use a shape adapted to the symbol orientation.

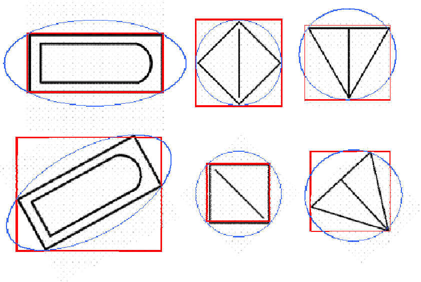
EPEIRES report 2005: Another shape to use is the parallelogram (rectangle without parallel edges). This object could have better overlapping properties than the ellipsis in the case of oriented symbols.

EPEIRES 2006: At the end we plan to use rotated bounding boxes, the boxes' centers (crossing point of the diagonals) will be use to locate the symbols.



Ernest: I agree, I think we must to keep things simple as possible for the groundtruth. I recommend here the oriented bounding boxes, that all.

Mathieu: We have to take care because the best shape not depends of symbols' orientations but of models. For some models the bounding box will be better in the no rotated case; for some others it will be the opposite. Also, for some models it will be impossible to find a best shape, all of them will give near overlapping rates. At the end to take into account all the possible cases you will have to consider a large number of different shapes (triangle, star, parallelogram, diamond, etc.).



3.2.2.2 Groundtruth is contours

Marçal: In order to avoid the weak precision problems of wrappers why not use contours (i.e. polylines) in the groundtruth.

EPEIRES report 2005: The key objective of our characterization is to detect the good and bad localizations (so to take a binary decision) more than having accuracy.

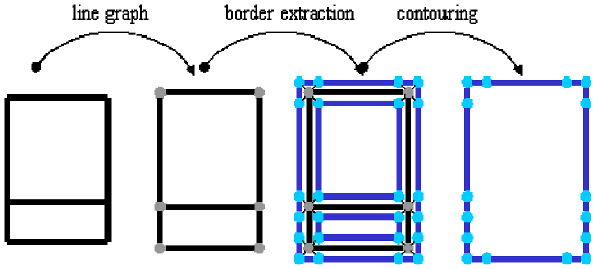
Mathieu: Yes, but you could have some limit cases for which you will take false decisions. To use polylines will allow to avoid these cases. You will compute ever better overlapping rates than using any bounding boxes, ellipses, parallelograms or others wrapper objects.

Hervé: Yes, but in the case of real-life documents to edit the contours in a manual way (from symbol to groundtruth) will be time expensive.

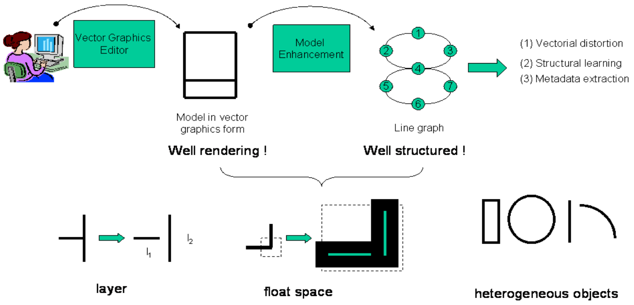
Mathieu: No, because we can keep the EPEIRES approach (i.e. to overlap the model images on the real drawings using vertical/horizontal scaling processes). We can next extract the external contours from model rasters in an automatic way. It exists many algorithms to do that like the one of [7]. This process is faster: it is done using a single line scanning step. However, I'm ok to say that these contours will less precise in comparison with the vector graphics ones.

Hervé: Extracting the contours from the vector graphics (circles, lines, arcs) could be also a complex operation, may be it will be harder to do.

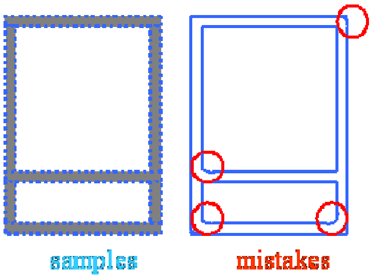
Mathieu: Yes you're right. This process consists in finding an object path (composed of arcs and/or straight lines) starting from a line graph. So, you have to build the line graph first. Once this line graph obtained it is easy to obtain the lines' borders, and then the external contours using techniques like [8] [9] [10].



Mathieu: The most complex part in this process is the building of line graphs [9]. Especially in our case because many mistakes exist in the symbol models as shown below. These mistakes are the unconnected ending points, the overlapped lines, the broken lines, the missing junctions, etc. Indeed, the models have been obtained using graphics editors (i.e. the WYSIWYG problem).



Mathieu: However, in order to strap this building problem, we can compute the contours from sample points of the lines' borders. Like this, it will be not necessary to build the line graphs. We will extract the sample points from all the lines' borders, and next from these points we will be able to compute the external contours. To reduce the complexity we can also reduce the number of point using a polygonalisation step. The obtained contours will be not so precise than with the previous method (they will contain some mistakes), but I suppose they will be enough fine for our evaluation objectives.



Marçal: Ok, you will loose precision but the mistakes are most of the time imperceptible. Moreover, if you compare the precisions gave by this approach with the wrappers so you have a deep gap.

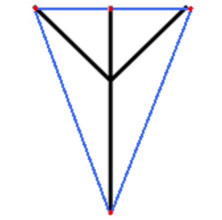
Mathieu: A last point to discuss concerns the complexity of this contour extraction step. I think it will be not a problem in all the cases. Indeed, we will have to compute the contours only one time during the whole evaluation process. Obviously, we can generate these contours from each model at the process beginning, and then to store them in the model file. Next, it will be only necessary to rotate or to scale these contours during the groundtruthing process.

Hervé: Comparing two contours (i.e. polygons) together could be also a complex operation.

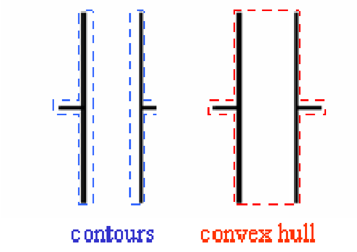
Mathieu: The complex part in this process is the junction detection between the two polygons. You will find a good tutorial here[2] on this topic. The most famous algorithm [11] do it with a O(n^2 log n) complexity (with n the number of segment). However, that we can do is to test the overlapping between the bounding boxes before starting the comparison of contours. The bounding box comparison is faster, this will limit the number of contour comparison to and then the whole complexity. An additional comment, I believe that our polygons won't be so big, so at the end the comparison time should be reasonable.

Marçal: You can use different techniques to strap the complexity problem. The first one is just to look for the outside / inside lines' points of polygon to compare. Another solution is to compare them at a raster level. I both cases you will approximate the overlapping, but you should loose few precision.

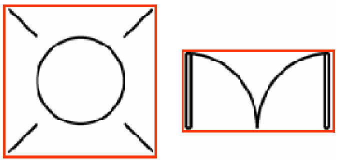
Mathieu: Other solution to reduce the complexity is to use "dominant contours". The key idea will be to look for the ending points of the lines, and next to find the convex hulls.



Marçal: I use convex hulls in my characterization approach. From my point of view, they represent a good complexity / accuracy deal between the wrapper objects and the "true contours". Another interesting feature of convex hulls is to represent the symbols with a single contour. Indeed, it is then easier to match two contours together than contour sets.



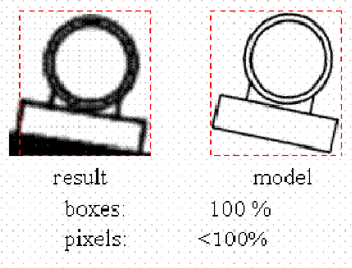
Mathieu: A problem still remains with the convex hulls, for some symbols the precision could be weak as with the bounding boxes. So, we gain precision for most of models but not for all. At the end this solution looks like wrappers, the precision gains depend too much of the considered model.



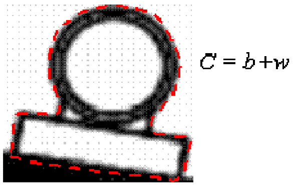
3.2.3 Groundtruth is rasters

Philippe: If we plan to use contours to be more precise, why not using directly the rasters ?

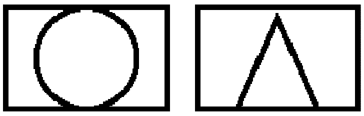
Mathieu: In our case we can't use directly the rasters. Indeed, if you compare at a pixel level you will be able to take into account elements of background. These elements will reduce then your overlapping rates as shown below. The only way to solve this problem is to use label maps to represent your data with three labels: black pixels of symbol, white pixels of symbol and other pixels that won't be considered (e.g. the borders).



Mathieu: Moreover, in both cases (contour and label map comparisons) you will obtain the same result. Indeed, when you compare the external contours of a given symbol, you compute then the *C* surface equals its number of black and white pixels.



Hervé: Yes, but two symbols could have a same external contour with a different pixel composition.



Mathieu: You're right, but for the localization we are not interested to have a similarity distance between the symbols (e.g.

[2] http://www.geometryalgorithms.com/Archive/algorithm_0108/algorithm_0108.htm

here a rate of common pixel) but just an overlapping criterion. So, in this case we need only the external contours to compute this overlapping rate. This means that the white and black pixels of symbols are considered as same in the case of overlapping.

Mathieu: Moreover, when you have models in a vector graphics format (like in the synthetic approach) you will lose precision in the contour comparison if you come-down to bitmap representation. Especially when you compare rotated symbols (the rotation distorts the raster models of symbols but not the vector graphics ones). Using contour of vector graphics will provide a better precision than using the ones extracted form the images. Also, using label maps needs more memory to store the objects to compare. It can also need more time computation according the maps' sizes.

Marçal: Ok, you will loose precision but the difference could be is very small. Concerning the come-down process I'm not sure that doing the raster comparison could need so more time.

3.2.4 Groundtruth is vector graphics

Jean-Yves: And what about vector graphics? the groundtruth could be an SVG or CAD representation of symbol or of the document we are looking for. It is our ultimate goal no?

Mathieu: In the case of the vectorization or the document understanding yes it is, but not in the case of the symbol spotting and recognition. We have just to provide labels or ranks with corresponding localization.

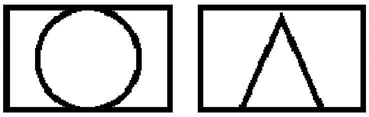
Mathieu: One additional precision, I think when Jean-Yves talks here about vector graphics he means set of unconnected lines. Indeed, "vector graphics" is an ambiguous term because you can store region objects (e.g. polylines, bounding boxes, ellipsis) into the vector graphics files. However, I haven't better word to propose, so we can keep but by taking care of the meaning confusion.

Mathieu: At this time it is the solution we have in the 3gT system (the synthetic document engine). We use models in a vector graphics format (SVG) from which we compute the bounding boxes.

Philippe: It is not the case in the EPEIRES platform: we use image models. When you deal with real documents you can't have models in a vector graphics form.

Marçal: There is a big confusion here, my felling is we still discuss about scanned paper documents. However, what about the digital ones (e.g. CAD and SVG files, screenshots)? In the true life I agree we have a large number of paper scanned documents, but also a huge part of digital ones. The evaluation framework has to deal with these digital documents. Having a SVG representation in the groundtruth is an excellent idea.

Mathieu: I think we have to take care to not do confusion between: to have a vector graphics groundtruth to evaluate the symbol recognition/spotting, or to evaluate the vectorization or document understanding. From my point of view using vector graphics in the groundtruth is more the purpose of the vectorization evaluation systems [25] [26]. We must keep in mind that the purpose of the localization characterization is not to have a similarity distance between the symbols (i.e. in this case it could be a rate of common lines) but just a surface overlapping. From the point of view of the localization the two following symbols are equals (if they are centered at the same point). Their similarity comparison will be done during the recognition and the ranking step (with the rank, the recognition label and the confidence rate). However having a vector graphics groundtruth allows to keep a full representation of symbols. I mean, if you store your groundtruth using bounding boxes you won't be able next to compute contours or convex hulls. In this case it could be a good idea to keep a full representation of symbols in the groundtruth.



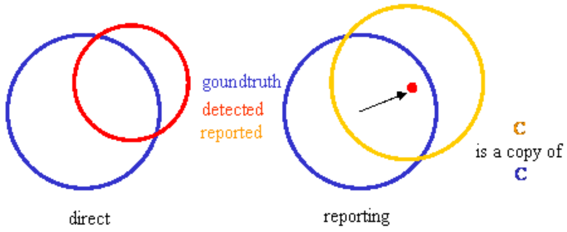
3.2.5 Results are points

Mathieu: When using points for the results a problem still remains, it will make the detection of merge cases (e.g. two groundtruthed objects recognized as a single one) impossible.

Jean-Yves: In the case of spotting systems this merge case occurs frequently, it must be taken into account and can't be considered as a missed one.

Marçal: I agree.

Mathieu: A solution to solve this problem (without using any region information from the systems) could be to use a reporting technique. In this case, we will use only the detected points by the systems. We will report next on each point the "ideals regions" defined in the groundtruth (i.e. bounding boxes, contours, etc.). Like this, we will be able next to detect the overlapping cases from point results.



Marçal: How will you deal with the orientations of symbols?

Mathieu: The orientation data are stored in the groundtruth. It will be then just necessary to translate the groundtruthed object of dx dy (the x,y distances between the gravity center of the groundtruthed object and the detected point of system).

Hervé: In an evaluation perspective of the symbol spotting the systems have to provide also region data. Indeed, final results will be presented as a sorted list of cropped images to users. So the systems must have segmentation results of spotted symbols. Using region result seems necessary if we want to evaluate the spotting.

Mathieu: I not agree because the results could be also viewed directly on the whole images. Each result could link a given localization point inside an image, the user could set the browser to adapt the view (to shift, to zoom).

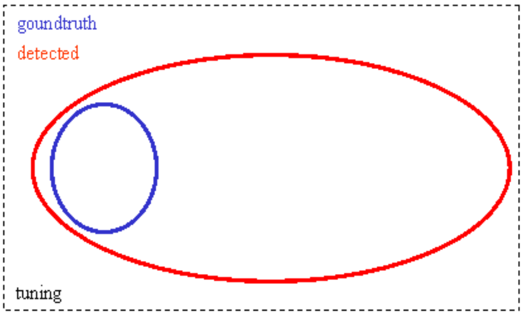
3.2.6 Results are regions

Mathieu: The first problem concerning this point is that not all the methods/systems will be able to generate region data.

Hervé: I agree. In this case I propose to make the evaluation of systems' regions as optional. I mean: to propose a characterization tool that will be able to take into account these regions, and next to let free the people to evaluate their systems on these aspects.

Marçal: It is fine for me if we compare all the systems at a same level. I mean, for the systems able to provide regions we have to compute the characterization as soon as from the points and from the regions. Not only from the regions because the systems can provide them.

Mathieu: In all the cases, to use the detected regions (i.e. results of systems) gives the possibility to the programmers to tune their systems in order to coerce the localizations. Indeed, they could be able to propose detected regions of large dimensions to constraint the overlapping with the groundtruth.



Marçal: I don't understand where is the problem in this case? This tuning is just a parameter that the participants have to precise. We shouldn't consider it as a cheat problem, if the participants can set their systems so it is a property of their approaches.

Mathieu: Yes, but a system could give very different results according to its setting. I don't want to forbid such settings, but may be it could be reasonable to fix a limit (e.g. not to provide Regions of Interest that will cover 80% of a drawing).

Hervé: May be one way to use the detected region and to solve this tuning problem is to use threshold for the overlapping (i.e. more than 50% then the symbols are overlapped).

Rashid: Well, as the visual analysis may vary from person to person (and system to system) it is very difficult to fix a common threshold. A resulting region may not include the whole symbol for a given system but let 80% of it, and to represent a good rate.

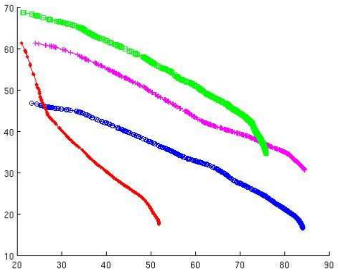
Mathieu: I agree, it is very difficult to use a common threshold because the dimensions of detected regions will change a lot according to the methods. At the end we could define a very dependent based threshold.

EPEIRES report 2005: The methods (statistical, structural or hybrids) could give very different results concerning the detected regions because they provide very different primitives. If we use their results to evaluate the localization we have to think about a very adaptable evaluation measure. The determination of a fixed threshold is not a good solution.

EPEIRES report 2005: To solve this problem one way could be to define several groundtruth (one by recognition method). However the groundtruthing step is already harder (i.e. it is time expensive for the real data the amount of data could be very high). So defining several groundtruth seem not a realistic way.

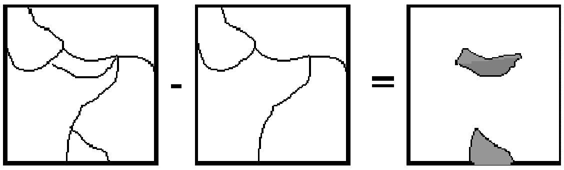
Hervé: To solve this problem we can use adaptable thresholding to the methods (i.e. not to have a binary decision step "rate > 0 then overlapped" but to do "for this method 30% corresponds to a large overlapping").

Marçal: I have another proposal. Let us to use deterministic overlapping criterion, and next to study the impact of the localization through the systems' results (i.e. the ROC curves). I present an example below. As you can see, the red method falls drastically whereas the three other perform similarly despite the difference in their ability to perform a fine or more coarse localization.



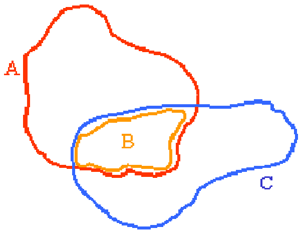
3.2.7 Mapping criterion

Mathieu: I propose to talk about mapping criterion concerning the evaluation of localization because it looks like a map comparison (i.e. any object set with localization data could be considered as a map) . We have to find where (and in which way) the maps are dissimilar.



Jean-Yves: To analyze the results it seems important to produce for each drawing a viewing of the result / groundtruth matching.

EPEIRES report 2006: The retained proposal to compute the global spotting criterion *s* for EPEIRES is:



$$Precision = \frac{B}{A} \quad Recall = \frac{B}{C}$$

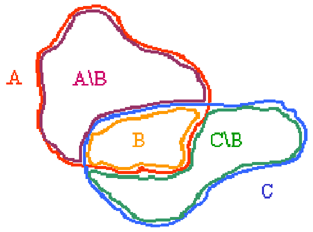
$$s = \frac{P+R}{2} = \frac{BA+BC}{2AC}$$

$$s^{*} = \frac{1}{\frac{1}{P}+\frac{1}{R}} = \frac{2B^{2}}{BA+BC}$$

EPEIRES report 2006: It exists perhaps several ways to compute such criterion. At the end the combination of the P and R is just a proposal, may be it will be more understandable to present the P and R in independent way.

Mathieu: Yes, like with a ROC curve.

Mathieu: From my point of view computing the mapping looks like a Precision / Recall computation but not exactly. I propose here another way to compute the rate *s* (the \ is the probability operator).



$$s=\frac{B}{A\setminus B+B+C/B}$$

Hervé: I have no precise idea on the problem but I more agree with the proposal of Mathieu than the one of EPEIRES.

Mathieu: In both cases the problem is how to evaluate the split and merge case (good or bad solutions). One way is to define parameter weights for all the cases and to set the system next (i.e. a split and merge case could be considered as 0.5 point, not a 1.0 point).

EPEIRES report 2006: The reject could be also considered for the localization evaluation.

3.3 Characterization of the matching

3.3.1 Characterization of the classification

Mathieu: Concerning the classification step there is "huge" works we can use from SymbolRec [12]. However what about the reject criterion of methods [13], there is no contribution in SymbolRec about that.

EPEIRES report 2006: The reject has to be considered in our evaluation scheme. It is important also to include other evaluation criteria (than the recognition rates) like the symbols' orientation and scaling.

Mathieu: Based on the learned data we have planed to produce (see section 2.3.2) it should not so complicate to do.

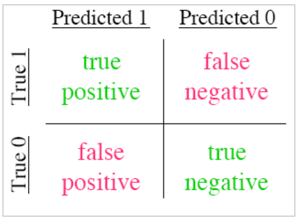
Marçal: It is necessary also to take into account the confidence rates of methods.

Mathieu: Yes but the distance used by the methods will be very different, how to compare that?

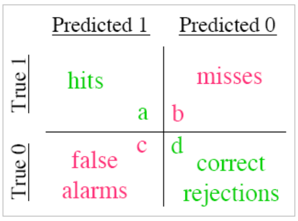
3.3.2 Characterization of the ranking

Mathieu: Concerning the spotting the systems have to provide some ranked lists with or without similarity measure. I think we can take benefit of the past experience of CBIR people with papers like [23].

Rashid: Another way for precision/recall accuracy can be based on the following confusion matrix [14]:



	Predicted 1	Predicted 0
True 1	true positive	false negative
True 0	false positive	true negative



	Predicted 1	Predicted 0
True 1	hits a	misses b
True 0	c false alarms	d correct rejections

3.4 System taxonomy

Jean-Yves: What other kind of "a priori" knowledge can be used to implement the systems, only the models of symbols, or such other information?

Mathieu: Pre-processing chains are an important feature we must take care.

Hervé: I suggest candidates should NOT use pre-processing of course, as they should not try to infer the background image.

Jean-Yves: I don't understand "candidates should NOT use" what does it mean?

Hervé: I mean that the evaluated systems should try to use the least knowledge on the domain. The purpose of them is not to separate components that can be discarded according to various criteria (width of the lines, length of the lines, and so on) from the initial set of components. We are not really interested on a criterion able to separate what belong to the symbol class from what do not belong to the symbol class (e.g. lets think about a document whom symbol layer is assigned to a specific color). Candidates should provide details on their methods. Tell if they think they take advantage of a specific preprocessing tool or previous knowledge domain.

Jean-Yves: I not agree with Hervé when says: "We are not really interested on a criterion able to separate what belong to the symbol class from what do not belong to the symbol class". In the contrary, I'm currently trying to find generic criteria able to separate what belong to the symbol class from other parts of the drawing in a generic way. For me it is the definition of what we call "symbol spotting".

Mathieu: From my point of view it seems difficult to coerce the candidates to use predefined knowledge set and pre-processing chains. That we can do is to fell free every body, but in the evaluation to subscribe the systems to given categories. These categories will be defined according to the used pre-processing and previous knowledge. Like this, we will compare the systems in partial way (with pre-processing or not, with learning databases or not, etc.). During the past editions of Contest the systems have been compared together without any distinctions. I suppose at this time it was a wrong way, like any competition you can't consider as same the participants (i.e. how to compare a moto-cycle with a cycle? the second uses no engine).

Jean-Yves: I agree with Mathieu and the last sentence of Hervé. The best solution is probably to let participants using what they want but they have to describe their method precisely (algorithms for preprocessing, knowledge used, etc.).

Marçal: I agree too.

Hervé: Of course, you're right. I wanted to highlight that any preprocessing tool is able to detect explicitly the background (or the text layer), and implicitly the symbol layer. Moreover, it will be a pity if such a treatment may be used without any knowledge of the symbols to be spot (this is the minimum) in the extreme case.

Mathieu: Ok, but the problem is now to define a method to compare the systems together (from a modelling point of view) that is not easy. Tony has done previous works on this topic [22].

Ernest: I agree, we have to define a model.

Jean-Yves: I'm agree too

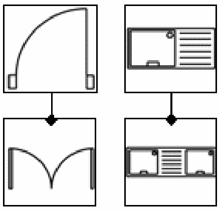
Hervé: Same.

3.5 Method profiling

Mathieu: Testing the scalability of methods is good thing, but I argue here that the usual evaluation process use the model sets (25, 50, 100, 150, etc.) in “black box” way. There is no previous idea concerning the types of symbol in these sets (are they composed of several or one connected component(s), did they contain arcs or/and straight lines, did they fill or not, from which domain did they belong 'electrical, architectural, etc.). Better than telling “this system works well with the symbol models 1, 16 and 32” it could be a great idea to have previous metadata concerning these models. The key objective could be to find “families” of model on which the systems work better. Some considerations are presented in [2] about this point, opposing the method vs. data based analysis of results. This topic raises the problem to extract and/or to define well formed metadata from symbol models.

Hervé: The analysis of results may take into account such metadata: does the symbol is disconnected or not? does it contains filled shapes? etc. I guess we have to produce as many characterization results as we have meta-classes of symbols.

EPEIRES report 2005: It exists some composition relations between some symbol models. May be having previous information concerning these compositions could explain some classification/ranking errors.



Jean-Yves: Yes, very good. But here we deal much more with recognition than with spotting, no? If necessary, in Tours we have databases of symbol model with filled shapes, mixed filled, thin shapes, flexible symbols, etc.

Mathieu: I'm not sure. The spotting uses a ranking step (i.e. similarity sorting) following the localization. The matching of systems' results with such metadata could be also applied to analyze the ranking. It could explain some errors detected in the precision/recall curve of some methods.

Ernest, Philippe: In the same way than the model scalability section (section 2.4.1) we can use metadata edited in a hand user way [24] to perform such characterization.

Mathieu: Concerning the profiling I'm not sure that this way is valid. Indeed, to find correlation between the systems' results and the symbols' metadata we will have to extract a large number of heterogeneous features from models. It seems difficult for me to do it without using an automatic approach.

4. References

[1] J. Lladós, E. Valveny, G. Sánchez, and E. Martí. Symbol recognition : Current advances and perspectives. In *Workshop on Graphics Recognition (GREC)*, volume 2390 of *Lecture Notes in Computer Science (LNCS)*, pages 104-127, 2002.

[2] K. Tombre, S. Tabbone, and P. Dosch. Musings on symbol recognition. In *Workshop on Graphics Recognition (GREC)*, volume 3926 of *Lecture Notes in Computer Science (LNCS)*, pages 23-34, 2005.

[3] E. Barbu, P. Héroux, S. Adam, and E. Trupin. Clustering document images using a bag of symbols representation. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1216-1220, 2005.

[4] S. Aksoy and al. Algorithm performance contest. In *International Conference on Pattern Recognition (ICPR)*, volume 4, pages 870-876, 2000.

[5] M. Delalandre, T. Pridmore, E. Valveny, H. Locteau and E. Trupin. Building Synthetic Graphical Documents for Performance Evaluation. Workshop on Graphics Recognition (GREC), pp 84-87, 2007.

[6] I. Phillips and A. Chhabra. Empirical performance evaluation of graphics recognition systems. *Pattern Analysis and Machine Intelligence (PAMI)*, 21(9):849-870, 1999.

[7] Y. Shima, T.Murakami, M. Koga, H. Yashiro and H. Fujisawa. A High Speed Algorithm for Propagation-Type Labeling Based on Block Sorting of Runs in Binary Images. In International Conference on Pattern Recognition (ICPR), pages 655–658, 1990.

[8] A. Ferreira, M. Fonseca, and J. Jorge. Polygon detection from a set of lines. In *Encontro Português de Computação Gráfica (EPCG)*, pages 59-162, 2003.

[9] D. M. Mount. Geometric intersection. In 2, editor, *The Handbook of Discrete and Computational Geometry*, pages 857-876, CRC Press, 2004.

[10] K. Zouba. Un algorithme pour la construction de graphes de régions à partir de graphiques vectoriels. Master's thesis, Laboratoire PSI, Université de Rouen, France, 2005.

[11] J.L.Bentley and T.Ottmann. Algorithms for reporting and counting geometric intersections. *Transactions on Computers (TC)*, 28(9):643-647, 1979.

[12] E. Valveny, S. Tabbone, O. Ramos, and E. Philippot. Performance characterization of shape descriptors for symbol representation. In *Workshop on Graphics Recognition (GREC)*, pages 82-83, 2007.

[13] T. C. W. Landgrebe, D. J. Tax, P. Paclik, and R. W. Duin. The interaction between classification and reject performance for distance-based reject-option classifiers. *Pattern Recognition Letters (PRL)*, 27(8):908-917, 2006.

[14] J. Davis and M. Goadrich. The relationship between Precision-Recall and ROC curves, 23rd international conference on Machine learning, Vol. 148, pp. 233 – 240, 2006.

[15] P. Dosch. Projet épeires, compte rendu de la réunion de démarrage. Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA), Nancy, France, 18 Mai 2005.

[16] P. Dosch. Projet épeires,compte rendu de réunion. Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes (LITIS), Rouen, France, 20 Janvier 2006.

[17] M. Rusinol and J. Llados. A region-based hashing approach for symbol spotting in technical documents. In *Workshop on Graphics Recognition (GREC)*, pages 41-42, 2007.

[18] H. Locteau, S. Adam, E. Trupin, J. Labiche, and P. Heroux. Symbol spotting using full visibility graph representation. In *Workshop on Graphics Recognition (GREC)*, pages 49-50, 2007.

[19] R. Qureshi, J. Ramel, D. Barret, and H. Cardot. Symbol spotting in graphical documents using graph representations. In *Workshop on Graphics Recognition (GREC)*, pages 39-40, 2007.

[20] S. Pearce and M. Ahmed. An evolutionary algorithm for general symbol segmentation. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 726-730, 2003.

[21] Y. Yu, A. Samal., and S. Seth. A system for recognizing a large

class of engineering drawing. *Pattern Analysis and Machine Intelligence (PAMI)*, 19(8):868-890, 1997.

[22] T. Pridmore, A. Darwish, and D. Elliman. Interpreting line drawing images: A knowledge level perspective. In *Workshop on Graphics Recognition (GREC)*, volume 2390 of *Lecture Notes in Computer Science (LNCS)*, pages 245-255, 2002.

[23] H. Muller, W. Muller, D. Squire, S. Marchand-Maillet, and T. Pun. Performance evaluation in content-based image retrieval: overview and proposals. *Pattern Recognition Letters (PRL)*, 22(5):593-601, 2001.

[24] E. Valveny and al. A general framework for the evaluation of symbol recognition methods. *International Journal on Document Analysis and Recognition (IJDAR)*, 1(9):59-74, 2007.

[25] I. Phillips and A. Chhabra. Empirical performance evaluation of graphics recognition systems. *Pattern Analysis and Machine Intelligence (PAMI)*, 21(9):849-870, 1999.

[26] L. Wenyin and D. Dori. Principles of constructing a performance evaluation protocol for graphics recognition algorithms. In *Performance Characterization and Evaluation of Computer Vision Algorithms*, pages 97-106. Springer Verlag Publisher, 1999.

[27] M. Fonseca, B. Barroso, P. Ribeiro, and J. Jorge. Retrieving vector graphics using sketches. In *Symposium on Smart Graphics (SG)*, volume 3031 of *Lecture Notes in Computer Science (LNCS)*, pages 66-76, 2004.

[28] M. Fonseca, A. Ferreira, and J. Jorge. Content-based retrieval of technical drawings. *International Journal of Computer Applications in Technology (IJCAT)*, 23(2-4):86-100, 2005.

[29] E. Valveny and E. Martí. A model for image generation and symbol recognition through the deformation of lineal shapes. *Pattern Recognition Letters (PRL)*, 24(15):2509-2907, 2003.

[30] M. Delalandre, B. Simon, S. Guillas, J.M. Ogier and K. Bertet. Straight Line Detection based on the Hough Transform a System and its Performance Evaluation. Draft, 2006.

[31] C. Ah-Soon & K. Tombre. Architectural Symbol Recognition Using a Network of Constraints. *Pattern Recognition Letters (PRL)*, vol (22), pp. 231-248 , 2001.