Recent contributions on the SESYD dataset for performance evaluation of symbol spotting systems

Mathieu Delalandre[1], Ernest Valveny[2], Jean-Yves Ramel[1]
[1]*Laboratory of Computer Science*
François Rabelais University, Tours city, France
firstname.lastname@univ-tours.fr
[1]*Computer Vision Center (CVC)*
UAB University, Bellaterra (Barcelona) city, Spain
ernest@cvc.uab.es

***Abstract*—In this paper, we present an overview about the use of the SESYD dataset for performance evaluation of symbol spotting systems. SESYD is a dataset of synthetic graphics documents containing non-isolated symbols in a real context. These documents are drawings (architectural floorplans and electrical diagrams), bags of symbols (i.e. arbitrary compositions of segmented symbols) or query symbols (i.e. cropped images of symbols). The whole dataset is currently composed of 5 document collections containing around 11,100 images representing 128,700 symbols. All these collections have been made publicly available for evaluation purpose since 2007. As a result, this dataset has been employed in a large number of papers in the literature. It constitutes till today, at the best of our knowledge, one of the top datasets in the graphics recognition community for performance evaluation. In this paper, we present an overview about the use of this dataset on the specific task of symbol spotting. We report and compare the main results and characterization approaches presented in the literature. We also present some improvements of our approach resulting in some new published collections.**

***Keywords*-symbol spotting, performance evaluation, distortion methods, performance characterization**

I. Introduction

Performance evaluation is a particular cross-disciplinary research field in a variety of domains (Information Retrieval, Computer Vision, CBIR[1], etc). Its purpose is to develop full frameworks in order to evaluate, to compare and to select the best-suited methods for a given application. Such a framework should include the groundtruth and datasets for training and testing, the definition of a data exchange protocol, the definition of metrics and the development of tools to match the system results with the groundtruth.

In the field of DIA[2], performance evaluation is a well known-topic since the early 90's. Performance evaluation frameworks have been defined for several DIA tasks such as table recognition, page segmentation, OCR[3], etc. In this paper we are interested in the performance evaluation of symbol spotting systems [1]. Since research on symbol spotting is just starting, it is still a little ambiguous to define "what a spotting method is". In [2], symbol spotting is defined as "a way to efficiently localize possible symbols and limit the computational complexity, without us ing full recognition methods". So, symbol spotting could be considered as a middle-line technique combining symbol recognition & segmentation and can also be viewed as a CBIR task.

As spotting is concerned by retrieval and localization of symbols, a hard problem is how to compare experimental results from existing systems. Traditionally, with recognition [3] this step was done independently for every system by comparing manually the results with the original images and checking the segmentation errors. This process was unreliable as it raises conflicts of interest and does not provide relevant results. Moreover, it does not allow to compare different systems and test them with large amounts of data. In order to solve this problem, several research works have been carried out over the last ten years on the performance evaluation of symbol recognition systems [1], resulting in the organization of several international contests and evaluation campaigns. However, these works have been focused on the recognition of isolated symbols. They do not take into account segmentation of symbols in real documents. One of the main reasons is the difficulty of obtaining a large set of documents with the corresponding groundtruth. Doing that manually would require an unaffordable amount of time, as all the symbols in the document must be precisely located and labeled. Groundtruthing still constitutes an open research problem today, and new approaches have been investigated recently on this topic [4].

In order to address this problem, we have proposed in [5] an approach to the generation of synthetic graphics documents containing non-isolated symbols in a real context. The initial version of the system has been published in 2007 [5] and a complete description of it was reported in [6]. Since 2007, this system has been used to generate a large variability of document collections containing graphics

[1]Content Based Image Retrieval
[2]Document Image Analysis
[3]Optical Character Recognition

and text information. These collections consist of drawings (architectural floorplan, electrical diagrams, geographical maps) but also of arbitrary compositions of segmented objects (e.g. bags of symbols, bags of words, cropped symbols, character images, etc.).

All these collections have been published in a dataset called SESYD[4] made publicly available[5] for performance evaluation purpose. This dataset is currently composed of 11 document collections containing around 280,000 images representing 440,000 objects (i.e. characters and symbols), whose 5 collections are concerned with the symbols (corresponding to around 11,100 images and 128,700 symbols). Since the publication of the initial collections in 2007, the SESYD dataset has been employed in around 30 papers[6] in the literature for different performance evaluation tasks including symbol recognition [?] & spotting [7], text segmentation [?], line drawing indexing [8], performance characterization [?], etc. It constitutes today, at the best of our knowledge, one of the top datasets in the graphics recognition community for performance evaluation.

In this paper, we present an overview about the use of this dataset focused on the symbol spotting task and the main results and conclusions reported in the literature. We also present some improvements of our approach and, as a result, some new published collections. In the rest of the paper, firstly we will present in section 2 a quick overview of our approach and give a presentation of the SESYD dataset including the new collections. Section 3 will report and compare the main results and characterization approaches presented in the literature. Finally, in section 4 we state the main conclusions about this work.

II. Previous work and recent contributions

A. Previous work

The system we have published in [5], [6] allows the generation of synthetic graphics documents containing non-isolated symbols in a real context. Our underlying aim was to make these documents as realistic as possible. However, realistic documents cannot be produced without taking into account human know-how into the process. In our work, we have considered an alternative approach to solve this problem. Our key idea came out observing that graphical documents are composed of two layers: a background layer and a symbolic one. We have used this property to build several document instances as shown in Fig. 1 (a). We generate several different symbolic layers and place them on the same background obtaining different documents. In this way, the building process of realistic documents can be considered as a problem of symbol positioning on a given document background.

[4]Systems Evaluation SYnthetic Documents

[5]http://mathieu.delalandre.free.fr/projects/sesyd/

[6]Statistics obtained with the "Publish or Perish" software.

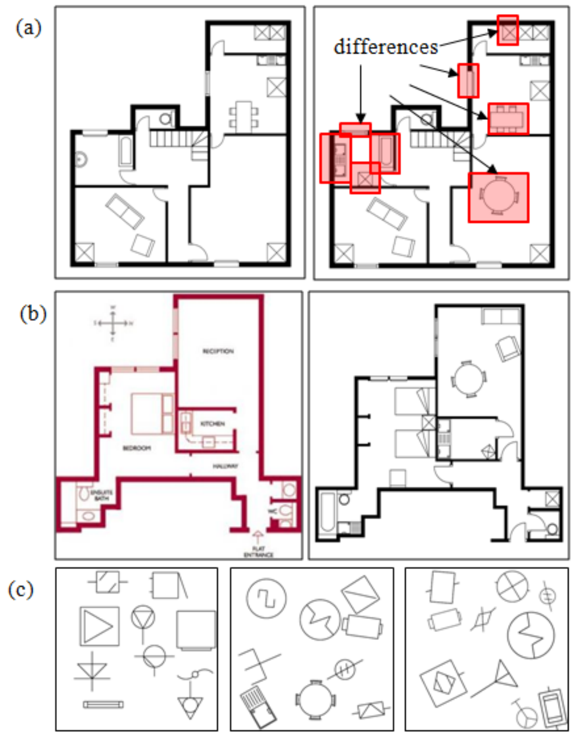

Figure 1. (a) two instances of document (b) reproduction of domain rules (c) bags of symbols

We have addressed the symbol positioning problem through the definition of sets of constraints. These constraints determine where and how the symbols can be placed on a background image according to the properties of a particular domain (architecture, electronics, engineering, etc.). They are edited with the help of users, familiar with the domains, to reproduce the rules of original drawings. The Fig. 1 (b) gives one example of an original floorplan and a corresponding synthetic document. The initial information concerning the types and the locations of symbols has been preserved in the constraints.

The Table I gives the details about the collections we have produced using our system in terms of numbers of datasets, images, symbols placed on the documents and symbol models. The datasets #1, #2 and #3 are the initial datasets presented in [6]. Our key objective was to highlight the flexibility and relevance of our approach for performance evaluation. The datasets #2, #3 are of drawings from the architectural and electronic domains, which proves the flexibility of our approach. In both datasets #2 and #3, the resolution of the produced images has been set in order to respect a mean symbol sizes of 320^2 and 288^2 pixels respectively. The dataset #1 is concerned with bags of symbols shown in the Fig. 1 (c). Here, the symbols are positioned at random on an empty background, without any connection, and using different rotation or scaling parameters. They establish a bridge between the recognition datasets published in the previous ISRC[7] and the localization datasets #2 and #3. In these bags, the symbols appear at a fixed size of 256^2 pixels. We have reported in [6] some performance evaluation experiments of a symbol localization system on these datasets. These experiments reflect variations in term of localization results between document instances, highlighting the relevance of our datasets for performance evaluation.

id	Collections	Datasets	Images	Symbols	Size	Models
#1	bags	16	1600	15046	256^2	25-150
#2	floorplans	10	1000	28065	320^2	16
#3	diagrams	10	1000	14100	288^2	21
#4	resolution	6×5	300×5	13106×5	320^2	16-21
#5	query symbols	6	6000	6000	288^2-320^2	16-21
		72	**11100**	**128741**		

Table I
COLLECTIONS OF TEST DOCUMENTS OF THE SESYD DATASET

B. Recent contributions

Recently, we have extended our dataset with a new collection #4 that consists of multi-resolution documents. Indeed, traditionally performance evaluation contests on symbol recognition [1] are focussed on binary degradation to distort images in a way similar to a scanning process. Degradation models, to be applied on the clean images resulting of the building process, have been proposed in the literature such as [**?**]. With this new collection #4, our goal is to investigate the impact of low-resolution as it noises the Web images. Fig. 2 (a) gives some examples of degradation resulting in modifications of the resolution. Our synthetic documents are initially generated in a vector graphics form with the corresponding groundtruth. To produce images at multiple resolutions, we apply a scaling process to the syntectic documents. The groundtruth is scaled simultaneously to preserve the corresponding graphics information. The obtained vector graphics documents are then rasterized in gray levels to produce the test images. The precise resolution of a rasterized vector graphics necessary for high-quality results depends on the viewing distance. Therefore the distortions will mainly result of this rasterization process. With a rasterization at low-resolution, the images will be blurred or pixelized excessively. We have selected 6 datasets from the collections #2 and #3 from the architectural and electrical domains. These datasets have been scaled to produce 5 resolution levels $1/2^n$, with $n \in \{0, 1, 2, 3, 4\}$. The Fig. 2 (a) gives examples of a cropped symbol at levels $1/1$, $1/2$ and $1/4$.

We have also extended our dataset with an additional collection #5 involved with distortions related to user interaction. This collection proposes query symbols (i.e. cropped

[7]International Symbol Recognition Contest, see [1] for references.

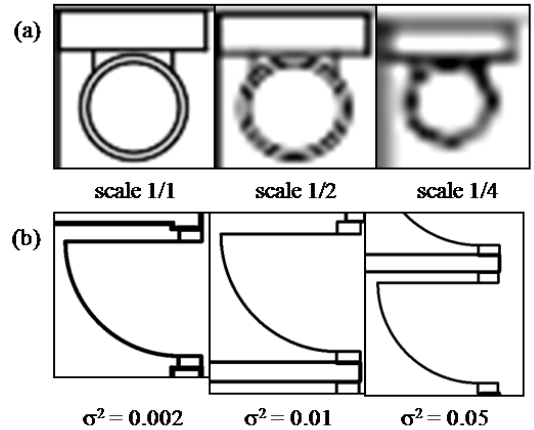

Figure 2. (a) Examples of degradation at different levels of resolution (b) Precision in cropping with different distortion parameters

images of symbols) extracted from complete drawings. The Fig. 2 (b) gives some examples of images extracted from this collection. These queries are symbols cropped from the images of drawings that can be affected by the way the user makes the selection. Then, this collection #5 tries to imitate this effect.

At the best of our knowledge, only the work of [**?**] deals with the topic of distortions related to user interaction in the literature. This work proposes a system that supports learning, from sample documents, of probabilistic models of objects which take into account their variability. These models can then be used to build and to distort objects. Focus distortions on real input provides more realistic test sets, but the work involved may be a significant proportion of that required to generate test data directly from real documents. In our approach we have adopted an opposed method which operates without any prior knowledge. While the proposed method does not correspond strictly to a “realistic” noise, we don’t need a learning step nor sample documents.

Our method is based on the generation of gaussian random numbers. We have employed here a gaussian model, as it is a common way to represent random distributions of real inputs. As shown in Fig. 3 (a), this generation consists of finding the v value from a normally distributed number s. The problem here is that the gaussian function cannot be inverted analytically. So in order to solve it the common way described in the literature is to use the *Box-Muller* transformation to pass from the uniform distribution to the gaussian one [9]. This transformation allows efficient computation of gaussian random numbers but does not allow the user to set different variance values σ^2, that can be necessary to control the amount of distortion (see Fig. 3 (b)). To solve this problem we compute an approximation of the gaussian sum by using the $erf(z)$ function. This function can be computed in various ways but it is usually computed as a Maclaurin polynomial. We need next to solve

$erf(z) \approx s$ in order to find the v value. This is difficult, because to obtain a good approximation of s, a high value of l (the order of the Maclaurin polynomial) is needed which involves a complex factoring step. In order to solve it we apply a dichotomic search algorithm on the x axis. Our experiments have shown that with a Maclaurin polynomial built with $l = 20$, the v values are obtained with 10^{-3} precision in less than 20 iterations. As the distortion process is done off-line, this processing charge has no impact.

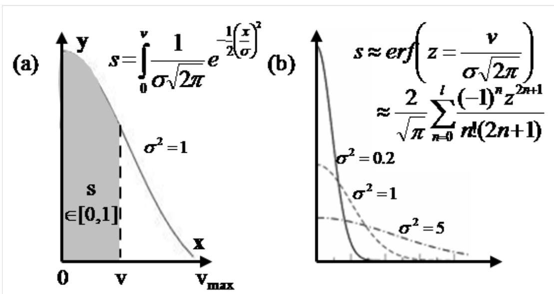

Figure 3. Generation of random gaussian numbers
(a) uniform distribution (b) no uniform distribution

We have used this random number generator within a distortion method to produce the query images. Our method starts from the groundtruth information about bounding boxes of symbols in drawings. We extract the top-left (x_0, y_0) and bottom-right (x_1, y_1) points of bounding boxes from the d_x, d_y values (i.e. width and height) as defined in Eq. (1). Next, we modify the coordinates of these points using four random gaussian numbers $v_{x0}, v_{y0}, v_{x1}, v_{y1}$ as detailed in Eq. (2) and (3). This process grows the bounding box around the symbol in order to simulate the over segmentation that usually appears during a crop of a user. The new coordinates $\{x_0^{'}, y_0^{'}, x_1^{'}, y_1^{'}\}$ will serve as cropping parameters to extract the query image from the complete drawing. We have used this approach to generate 6 datasets of query symbols (Table I) from the architectural and electrical domains, using three different settings for the random number generator $\sigma^2 = \{0.002, 0.01, 0.05\}$. Fig. 2 (b) gives examples of query symbols produced at these three levels of distortion.

$$x_0, y_0 \qquad d_x, d_y \qquad x_1 = x_0 + d_x, y_1 = y_0 + d_y. \tag{1}$$

$$x_0^{'} = x_0 - v_{x0} \times d_x \qquad y_0^{'} = y_0 - v_{y0} \times d_y \tag{2}$$

$$x_1^{'} = x_1 + v_{x1} \times d_x \qquad y_1^{'} = y_1 + v_{y1} \times d_y \tag{3}$$

III. Report of use

This section reports the use of the SESYD dataset in the literature. As stated in introduction, since its initial publication this dataset has been employed in around 30 papers in the literature. We present here an overview about the use of this dataset focussed on the symbol spotting task, and next we report and compare the main results and characterization approaches presented in the literature. Since its publication, the SESYD dataset has attracted a strong interest in the research community on symbol spotting systems [**?**], [7], [10], [11], [12], [13][8]. These systems are focussed on symbol retrieval in complete drawings, using as input query symbols produced by users. The important characteristics for this kind of systems are scalability, according to different criteria like application domains or size of the database to index, robustness to noise, etc. Compared to the traditional recognition methods no training datasets are used here and the complexity of methods is a major issue to support a fluent user interaction. The Table II - left part allows a comparison of the main results reported in [**?**], [7], [10], [11], [12], [13]. As these works are focused on symbol retrieval in complete drawings, the collections #2 and #3 were mainly concerned. The input of systems are sets of query symbols M/Q (being M the number of different models and Q the total number of query symbols) to be retrieved in the collection of drawing images $Datasets$. Results are presented in terms of (P)recision/(R)ecall metrics that are well known in the Information Retrieval field.

	Dataset	Symbols	M/Q	P/R	1) Size	2) Models	3) Domains	4) Queries	5) Protocol	6) Complexity
[14]	#2(15)	205	8/8	1.0/0.69	-	-	+	-	-	NP
[15]	#2(15)	220	9/9	Na/0.87	-	-	+	-	-	Linear
[7]	#2(300) #3(20)	12513 223	8/8 14/14	0.97/0.98 0.94/0.98	+	+	++	-	++	Exp
[16]	#2(50) #3(36)	Na	10/10 4/4	0.95/0.90	+	-	++	-	-	Linear
[6]	#2(200) #3(800)	9600 4216	21/1000 16/1000	0.56/0.90 0.81/0.90	++	++	++	+	+	Linear
[17]	#2(1000)	28065	16/16	0.79/0.93	++	++	+	-	++	Sub-linear

Table II
Comparison of symbol spotting systems

$\#n(m)$ m images of the collection n, M/Q are (M)odels and (Q)ueries
P/R are (P)recision and (R)ecall, (-,+,++) are comparative scores
Na is (N)o (a)vailable

In the rest of the section we give a detailed analysis of the performance evaluation approaches reported in these papers and, as a conclusion, a more objective comparison of spotting results of these systems. We will argue this comparison on several aspects including system scalability (tests, models and domains), characterization protocols (query symbols, results) and processing complexity. The table II - right part gives comparative scores we propose

[8] Among the existing publications of the same authors, we have selected the most recent and relevant ones.

for each of these systems on these aspects. We will present each of the aspects in the next subsections 1. to 6., and discuss and justify the scores. At last, we will compare and discuss the spotting results of the systems in the subsection 7.

1) Scalability according to the size of the test database: To drive the experiments, the authors have extracted subsets from the collections #2 and #3 initially constituted of one thousand drawings each. Some of these experiments concern very small subsets. In [10] and [11] the authors provide results on a dataset of 15 drawings composed of around 200 symbols[9]. The rest of the experiments deal with with hundreds of drawings when possible, 86 in [12] and 320 in [7]. The top experiments in terms of size have been reported in [**?**] and [13], with one thousand drawings and around 14,000 and 28,065 symbols respectively.

2) Scalability according to the number of symbol models: One of the main limitation of the proposed methods is scalability of symbol models. The collections #2 and #3 have been produced from symbol model libraries of size 16 and 21 respectively. However, most of the proposed systems [10], [11], [7], [12] have restricted their experiments on subsets of these models, (8-10)/16 for the floorplans and (4-14)/21 for the electrical diagrams. Only the works described in [**?**], [13] address the spotting task using the complete model libraries. Few explanation are given in the corresponding papers about this problem, but one of the main reasons is certainly the descriptive power of the employed descriptors. In [10], [11], the authors use a graph based representation describing closed loops detected on the drawings with their adjacency relations. However, not all the symbols are composed of loops. The works proposed in [7] and [12] exploit local descriptors (i.e. key points) to be used in matching processes focussed on mapping transformation [7] and Shape Context [12]. The major problem in these approaches is the robustness of the key point detection step. Sampled points of contours are used in [7], introducing occlusions when symbols appear connected. In [12] the DoG[10] descriptor is used. This descriptor is popular in the computer vision field, it focusses on describing the local structures of images based on intensities, orientation and location histogram. However, its performance is likely to be reduced when facing to bi-level document images.

3) Scalability according to the domain usability: Domain scalability concerns the ability of systems to process drawings of different domains. This aspect is strongly linked to the scalability of symbol models, as restrictions on models will impact the domains concerned by these models. Most of the systems [**?**], [7], [12] have been tested on drawings from the architectural and electrical domains using the #2 and #3 collections. As stated before, some of them [7], [12] include limitations on the processed symbol models: (8-10)/16 for the floorplans and (4-14)/21 for the electrical diagrams. Results are given in separate way to prove the domain adaptation of the methods [7] [**?**]. The work described in [12] is the only one combining the results of the both domains within the same experiment.

Only the works described in [10], [11], [13] report results on the architectural domain only. The works [10], [11] use a graph based representation describing closed loops detected on the drawings with their adjacency relations. A loop based representation seems to be little appropriate for electrical symbols, as symbols mainly appear as sets of parallel and orthogonal straight lines without loops (e.g. capacitor, earth, resistor, etc.). A similar constraint occurs in [13]. In their approach, the authors represent the symbols with graph paths. These paths are computed from a line graph resulting of a vectorization process. Next, they are described using some statistical descriptors in order to be recognized in a classification chain. The approach seems well adapted for the strong connected symbols as those in architectural floorplans, however the path based representation looks limited when the symbol representation is driven by the neighboring information only e.g. parallel lines of capacitor or earth symbols.

4) Types of query symbols: The spotting systems are focussed on symbol retrieval in complete drawings, using as input query symbols produced by users. Consequently, a set of query symbols must be defined and submitted to the systems to evaluate their P/R abilities. These query symbols should present distortions as they are supposed to be provided by users. However, most of the proposed systems [11], [7], [10], [12], [13] have restricted their query symbols to the ideal case only. Symbol model images have been used resulting in clean and unoised query symbols, that is far away from a real spotting use-case. As a result, characterization have been computed from too small amounts of query symbols (8-22) to really judge about the performances of the systems. Only the work done in [**?**] has considered a complete evaluation. The authors have exploited the collection #5 of query symbols to evaluate their spotting approach. At the end, spotting has been tested from two thousands query symbols at the first level of distortion (i.e. $\sigma^2 = 0.002$) to retrieve the 14,000 symbols. No results are presented about the highest levels of distortion, that could certainly impact the P/R results.

5) Characterization protocols: In the papers [7], [10], [11], [12], [**?**], [13] the characterization of results is given with the P/R metrics. Indeed, as the symbol spotting looks like a CBIR task, the P/R seem to be well adapted. However, symbol spotting concerns not only the retrieval of the images containing the query symbols, but also the localization of the symbols in the whole images of drawings. Indeed, the symbols appear connected in a drawing, and each drawing is usually composed of several symbols. Therefore

[9][11] is an improvement of [10], datasets and features are the same in both experiments.

[10]Difference-of-Gaussian

the characterization metrics should take into account the retrieval and localization simultaneously. Some works have been published recently to propose some adapted characterization protocols for symbol spotting and localization [?], [14]. Despite these contributions, this topic still constitutes today an open problem as no comparison of the protocols has been proposed in the literature, and no ready-to-use characterization tools are available.

As a result, the works described in [?], [7], [10], [11], [12] have defined their own characterization protocols. In [10], [11], [12], the characterization has been done by comparing the results with the original images and checking manually the spotting errors. This process was unreliable as it raises conflicts of interest and does not provide relevant results. In addition, doing the comparison manually is a strong limitation for the scalability of tests, as the error checking becomes time consuming. The authors in [?] have chosen an automatic characterization protocol that doesn't take into account the results of localization. In their protocol, a detection is considered as true positive if a spotted symbol in a drawing appears at least one time, at any localization, in the corresponding groundtruth file.

Only the works described in [7], [13] exploit the localization of symbols for performance characterization. In [7], a characterization protocol is proposed exploiting the bounding boxes of the spotted symbols to be compared with those of the groundtruth. A symbol is considered as well detected if no more than 10% of overlapping is detected with the groundtruth. This protocol looks like the one of [?] well known for the characterization of the logos spotting systems. The authors in [13] use the protocol described in [?]. This protocol is slightly different from the previous one [7] as the characterizations, at retrieval and localization levels, are combined in a single metric. This metric combines the precision (P) and recall (R) metrics with areas of the overlapping between sets of polygons representing results and ground-truth. Final characterization is given by the way of an average precision which rewards the earliest return of relevant items. In [?], the authors present this metric as especially adapted to evaluate well the behavior of symbol spotting systems.

6) Complexity: In a retrieval application, the complexity of methods is a major issue to support a fluent user interaction. This consideration is quite accepted today in the graphics recognition community to define what a symbol spotting system should be [1]. However, few of the works published in [10], [11], [7], [12], [?], [13] report time results and a complexity analysis of their approach. In table II, we propose a classification of these different systems regarding the decreasing levels of complexity NP[11] & exponential, and linear & sublinear.

NP & exponential complexities: In [10], spotting is achieved using a subgraph isomorphism algorithm. The subgraph isomorphism is of NP complexity. However the employed matching algorithm in [10] is tuned for large graphs using heuristics resulting in approximation of the isomorphism. They argue that the matching is NP bounded in worst of the cases. Despite this optimization, it seems difficult to apply this approach to a real spotting use-case. The authors in [7] propose an approach where spotting is done by applying a branch-and-bound algorithm using feature points extracted from drawings. The worst case complexity of any branch-and-bound algorithm is exponential but the average observed complexity is significantly lower. The average running time reported in the paper is $22s$ per query, that restricts this system to off-line uses in most of the cases.

Linear & sub-linear complexities: At the best of our knowledge, only the works [?], [11], [12], [13] have considered the complexity problem in their systems. In [11], the approach of [10] is extended on the complexity aspects as they have shifted the subgraph isomorphism problem to a linear integer programming. Another work interested with replacement of subgraph isomorphism algorithms due to complexity is [?]. Their method accomplishes subgraph spotting through graph embedding working within a linear time complexity. In [13], the complexity is addressed using indexing data structure to spot symbols in a sub-linear time complexity. Their approach relies on hash tables to organize their shape descriptors. They employ a specific hashing data structure, called locality sensitive hashing (LSH), that aims to organize the shape descriptors regarding the neighborhood information. Experiments done report an average time retrial of around 0.1 second per query corresponding to the shortest retrieval time reported in the literature.

7) Comparaison of the results: All the works described in [?], [7], [10], [11], [12], [13] present results, in terms of P/R, to characterize the spotting ability of their systems. As these results have been obtained from the SESYD dataset, this could open interesting perspectives for comparison of these different systems. Besides, some of these works [7], [12] present comparisons of their results with other systems argued by the use of this common dataset for their experiments. However, as the characterization approaches employed in these works differ on many aspects (i.e. used instances of the SESYD dataset, symbol models supported by the spotting systems, types of query symbols to be submitted, characterization protocols, etc.) a direct comparison of the P/R results remains quite subjective and must be considered carefully.

The results presented in [10], [11], [12] have been obtained without automatic characterization process and from very small instances of the SESYD dataset. As a conclusion, results presented in these works seem few reliable and it becomes difficult to give relevant conclusions about the evaluation of these systems.

[11]Nondeterministic Polynomial

Among the exiting works interested with the SESYD dataset, the main relevant results from our point of view have been reported in [?], [7], [13]. The results presented [?] present a particular property, as they have been obtained from a large variability of query (in terms of noise and number) to be submitted to the system. At the end, spotting has been tested from two thousands query symbols with a low level of distortion. The dataset used to obtain these query symbols is the collection #5 presented as new contribution in this paper. However, as the characterization has been driven without taking into account the localization level, the presented P/R scores certainly overvalue a lot the real performances of the system. The authors should also complete their experiments with the highest levels of distortion, that will certainly impact significatively the spotting results. In [7], [13], complete results have been presented including characterization at the retrieval and localization levels. As the characterization metrics and symbol librairies differ in these both works, none direct comparison of the P/R scores can be driven objectively. However, results reported in these papers prove than, considering the case without distortion of the drawing images and the query symbols, retrieval can be achieved with a quite good recall and acceptable precision. Let's notice than improvements must be driven for the system [7] on the scalability aspects, as not all the symbol models are supported to date. In addition, the proposed approach should also be revisited on the complexity of the proposed approach, as the reported retrieval times cannot guaranty on-line uses in most of the cases.

As a conclusion, we have proposed in this section a quite detailed analysis of the results and performance evaluation approaches reported in these woks. This analysis highlights some of the strong and weak points of the proposed approaches. As a conclusion of this analysis, we give here a set of recommandations than can be used as guideline for performance evaluation of further researches on the symbol spotting systems.

1) The characterization process must be achieved automatically and computed from a significant number of test documents to be relevant.
2) The experiments must be done using the full symbol model library to allow an exhaustive comparison of the systems.
3) The types of query symbols to be submitted to the systems must be extended in terms of noise and number, as query symbols should present distortions introduced by users.
4) The characterization protocols must take into account the retrieval and localization levels, as spotting concerns the localization of the symbols in the whole images of drawings.
5) The complexity of methods is a major issue to support a fluent user interaction, then spotting must be considered as a real-time application where results provided after the deadline are not only in late but wrong.

IV. Conclusions

In this paper we have presented overview about the use of the SESYD dataset for performance evaluation of symbol spotting systems. SESYD is a dataset of synthetic graphics documents that has been used in a large number of papers in the literature for different performance evaluation tasks (symbol recognition & spotting, text segmentation, line drawing indexing, etc.). It constitutes today, at the best of our knowledge, one of the top datasets in the graphics recognition community for performance evaluation. Considering the documents related to the symbol spotting task, the SESYD dataset is currently composed of 5 document collections containing around 11,100 images representing around 128,700 symbols. These documents contain non-isolated symbols in a real context, including drawings, bags of symbols or query symbols.

In this paper, we have presented a quick overview of our approach with some improvements resulting in new published collections. A first new collection #4 consists of multi-resolution documents, with the goal to investigate the impact of low-resolution as it noises the Web images. We have also proposed an additional collection #5 of query symbols (i.e. cropped images of symbols) extracted from complete drawings. These query symbols tries to reflect the way the user makes the selection of symbols on drawings. In our approach we have adopted a method which operates without any prior knowledge, based on the generation of gaussian random numbers. While the proposed method does not correspond strictly to a "realistic" noise, we don't need a learning step nor sample documents.

In a second part, we have reported and compared the main results and characterization approaches for performance evaluation of symbol spotting systems presented in the literature [?], [7], [10], [11], [12], [13]. Despite the common use of our dataset for performance characterization in all these works, a direct comparison of the system results remains quite subjective and must be considered carefully as the employed characterization approaches differ on many aspects. However, results reported in the some papers [7], [13] prove than, considering the case without distortion of the drawing images and the query symbols, retrieval can be achieved with a quite good recall and acceptable precision. We have proposed here a quite detailed analysis of the results and performance evaluation approaches reported in these woks, resulting in some main recommandations than can be used as guideline for performance evaluation of further researches on the symbol spotting systems.

To conclude, we are enjoyed today of the interest met in the research community on the SESYD dataset for performance evaluation of the symbol spotting systems. As this dataset has attracted many researches, we believe it constitutes a useful contribution. Results reported in the literature highlight quite good performance on the first collections #2 #3. In that way, we can consider that researches on symbol spotting get through a first stage with the help of this dataset. However, these collections #2 #3 are not concerned by distortion. In addition, the used model libraries are restricted in terms of size, however real domain applications usually concern thousands of symbol models. Thus, a huge work is waiting the research community on symbol spotting, as researches on this topic should go ahead on the distortion and scalability aspects. We believe than the new collections #4 #5 we have presented in this paper will help to initiate work in this direction.

V. Acknowledgement

The authors wish to thank Partha Pratim Roy (Laboratory of Computer Science, François Rabelais University, Tours, France) for his comments and corrections about the paper.

References

[1] M. Delalandre, E. Valveny, and J. Lladós, "Performance evaluation of symbol recognition and spotting systems: An overview," in *Workshop on Document Analysis Systems (DAS)*, 2008, pp. 497–505.

[2] K. Tombre, S. Tabbone, and P. Dosch, "Musings on symbol recognition," in *Workshop on Graphics Recognition (GREC)*, ser. Lecture Notes in Computer Science (LNCS), vol. 3926, 2005, pp. 23–34.

[3] J. Lladós, E. Valveny, G. Sánchez, and E. Martí, "Symbol recognition : Current advances and perspectives," in *Workshop on Graphics Recognition (GREC)*, ser. Lecture Notes in Computer Science (LNCS), vol. 2390, 2002, pp. 104–127.

[4] M. Delalandre, J. Ramel, and N. Sidere, "A semi-automatic groundtruthing framework for performance evaluation of symbol recognition & spotting systems," in *Workshop on Graphics Recognition (GREC)*, 2011.

[5] M. Delalandre, T. Pridmore, E. Valveny, E. Trupin, and H. Locteau, "Building synthetic graphical documents for performance evaluation," in *Workshop on Graphics Recognition (GREC)*, 2007, pp. 84–87.

[6] M. Delalandre, E. Valveny, T. Pridmore, and D. Karatzas, "Generation of synthetic documents for performance evaluation of symbol recognition & spotting systems," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 13, no. 3, pp. 187–207, 2010.

[7] N. Nayef and T. Breuel, "Statistical grouping for segmenting symbols parts from line drawings," in *International Conference on Doucment Analysis and Recognition (ICDAR)*, 2011.

[8] ——, "Building a symbol library from technical drawings by identifying repeating patterns," in *Workshop on Graphics Recognition (GREC)*, 2011.

[9] W. Janke, "Pseudo random numbers: Generation and quality checks," in *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*, ser. Nic Series, vol. 10, 2002, pp. 447–458.

[10] H. Locteau, "Contributions a la localisation de symboles dans les documents graphiques," Ph.D. dissertation, Université de Rouen, Rouen, France, 2008.

[11] P. L. Bodic, H. Locteau, S. Adam, P. Heroux, Y. Lecourtier, and A. Knippel, "Symbol detection using region adjacency graphs and integer linear programming," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2009, pp. 1320–1324.

[12] T. Nguyen, "Localisation de symboles dans les documents graphiques," Ph.D. dissertation, Université de Nancy 2, Nancy, France, 2010.

[13] A. Dutta, J. Llados, and U. Pal, "Symbol spotting in line drawings through graph paths hashing," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2011.

[14] M. Delalandre, J. Ramel, E. Valveny, and M. Luqman, "A performance characterization algorithm for symbol localization," in *Workshop on Graphics Recognition (GREC)*, ser. Lecture Notes in Computer Science (LNCS), vol. 6020, 2010, pp. 260–271.