Symbol descriptor based on shape context and vector model of information retrieval

Thi Oanh Nguyen, Salvatore Tabbone, Oriol Ramos Terrades

Symbol descriptor based on shape context and vector model of information retrieval

T-O Nguyen, S. Tabbone, O. Ramos Terrades
LORIA - Université Nancy 2
Campus Scientifique - BP 239
54506 Vandoeuvre-les-Nancy, France
{thi-oanh.nguyen,tabbone,ramos}@loria.fr

Abstract

In this paper we present an adaptive method for graphic symbol representation based on shape contexts. The proposed descriptor is invariant under classical geometric transforms (rotation, scale) and based on interest points. To reduce the complexity of matching a symbol to a large set of candidates we use the popular vector model for information retrieval. In this way, on the set of shape descriptors we build a visual vocabulary where each symbol is retrieved on visual words. Experimental results on complex and occluded symbols show that the approach is very promising.

1. Introduction

Symbol retrieval in technical documents is still a hot challenge in the document analysis community and shape representation for symbol recognition has been the subject of much research. Extensive surveys of shape analysis can be found in [6, 9]. The choice of a particular representation scheme is usually driven by the need to cope with requirements such as robustness against noise, stability with respect to small distortions, invariance to common geometrical transformations or tolerance to occlusions. Usually, two classes of feature descriptors are encountered: those that work on a shape as a whole (called region-based descriptors) and those that work on the contours of the shape (called contour-based descriptors). Usual contour-based descriptors include Fourier descriptors [13, 21] which have been widely used. Region-based descriptors take into account all the pixels within a shape and common methods are based on moment theory [3, 12].

To retrieve similar objects a measure of similarity is defined between feature descriptors which measures the distance between them. However, a similarity comparison is not often simple and more transformations are required to achieve that. Moreover, the size of the descriptor is often high with redundant information and this introduces a high complexity when searching similar objects into large collections of documents.

This paper is a part of an ongoing work tackling this problem. It is based on previous works of shape recognition and image retrieval [4, 16]. The overall approach is outlined in figure 1. First, we introduce a Shape Context descriptor computed on Interest Points (SCIP). We have adapted the shape context [4]. More precisely, we extract on each symbol interest points and, on a neighbourhood (a local context) of these points, we compute a descriptor. The use of shape descriptor is motivated by the nature of documents which are most often in grey scale, or in binary. Hence, a shape descriptor is well-suited to capture information in such documents. We have chosen shape context for its performance on partially occluded objects [4] and symbols appear, after a segmentation step, also partially occluded when they are embedded into a graphical document.

Secondly, shape context descriptors are clustered to build a visual vocabulary which is a kind of abstraction process. Each centroid cluster is considered as a visual word and shape context descriptors in the same cluster share similar shape information, regardless the symbol where the point of interest has been extracted. Finally, each symbol is described by visual words and matched against a symbol query. The approach is similar to [16] where an efficient retrieval is achieved using inverted files based on text vector model and frequency weightings.

This paper is organized as follows. In section 2, we describe an adapative method for graphic symbol representation based on shape contexts. The symbol retrieval system using the vector model is introduced in section 3. In section 4, we present the adaptability of our method for graphic symbols. Experimental evaluations on the GREC database are given. Finally, we conclude and give perspectives to our work (section 5).

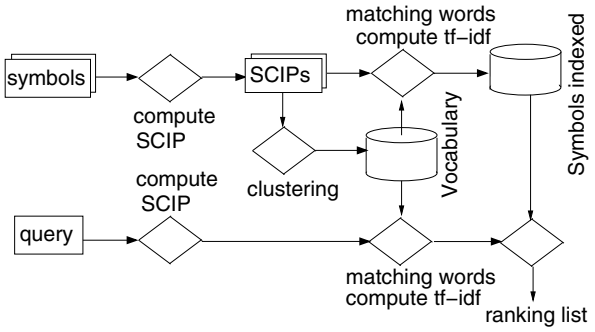

Figure 1. Symbol retrieval system

2. Shape Context for Interest Points (SCIP)

In this section, we present an adaptive solution for graphic symbol representation based on shape contexts. The computed descriptors are invariant under rotation and scaling. As shape contexts contain rich information about the local geometry of object, the proposed solution allows us to keep only useful information.

2.1. Shape context recall

The *shape context* of a point p_i belonging to the contour of an object is determined by the distribution of contour points in the surroundings of p_i [4],. It is a bivariate histogram h_i of the relative coordinates of the contour points.

A shape is represented by a discrete set of points sampled from its internal and external contours $\mathcal{C} = \{p_1, p_2, .., p_n\}, p_i \in \mathbb{R}^2$ where *n* is the number of contour points on the shape. For a point p_i, the relative coordinates of remaining *n-1* points are calculated to build its histogram h_i. These relative coordinates are the coordinates of points in log-polar coordinate system using p_i as the origin.

$$q = (logr_q, \theta_q), \forall q \neq p_i \wedge q \in \mathcal{C} \quad (1)$$

where r_q is the distance from q to p_i and, θ_q is the angle formed between vector $\overrightarrow{p_i q}$ and horizontal axis. The shape context (h_i) of point p_i is defined by (2):

$$h_i(l) = \#\{q \neq p_i : (q - p_i) \in bin(l)\}, l = \overline{1, L} \quad (2)$$

$h_i(l)$ is the number of contour points in the l^{th} bin $bin(l)$. Therefore, an object $\mathcal{O}$ is described as a set of shape contexts associated with the contour points.

$$\mathcal{O} \equiv \{h_i | p_i \in \mathcal{C}\} \quad (3)$$

However, the shape context described above is not invariant under scaling and rotation. To achieve scale invariance, all radial distances are normalized by the mean distance α between the n^2 point pairs in the shape [4]. For rotation invariance, the authors proposed to use the tangent vector at each point as the positive x-axis instead of absolute axis for computing the associated shape context.

2.2. Shape context vs. SCIP

Shape context is an extremely rich descriptor, the shape context of a point contains important information in the surrounding. In this perspective, the description of an object using shape contexts of all contour points represents a big set with redundant elements. However, there are many studies that show that an object can be efficiently detected from its keypoints [1, 5, 8, 10, 16]. Hence, in our context, we retain only the shape contexts of characteristic points in the symbols known as interest points.

Interest point detection: many methods for detecting interest points have been proposed [14, 18]. We have chosen DoG (Difference-of-Gaussian) keypoints detector that is introduced in [8] for our experiments though other detectors (Harris-Laplace, Hessan-Laplace, ...) are possible. The interest points in an image are considered as the extrema in a scale-space pyramid built with DoG filters (see (4)). In the evaluation by Mikolajczyk and al. [11], SIFT (Scale Invariant Feature Transform) descriptor calculated at keypoints detected by DoG detector outperforms others. In addition, as the DoG operator is a close approximation of the Laplacian-of-Gaussian function, consequently, most of the detected points are nearby the junctions of object model [18] which play a important role in distinguishing one model from another.

$$D(x, y, \delta) = (G(x, y, k\delta) - G(x, y, \delta)) * I(x, y) \quad (4)$$

Computing shape contexts of interest points (SCIP): Now, suppose $\mathcal{IP} = \{P_1, P_2, .., P_N\}$ is the set of interest points and $\mathcal{C} = \{q_1, q_2, .., q_n\}$ is the set containing contour points of object. Each point in $\mathcal{IP}$ is considered as the reference point to compute its shape context. We would like that the descriptor is invariant under scaling and rotation, thus the relative coordinates of contour points must be normalized. However, the interest points are rarely contour points [18] eg. $\mathcal{IP} \nsubseteq \mathcal{C}$, the tangent vector (as proposed in [4]) is not a practical parameter used for normalizing. Instead of using it, we choose the dominant orientation of interest point as the positive x-axis. Therefore, each interest point P_i is represented by its coordinates and the dominant orientation is:

$$P_i = \{x_i, y_i, \vec{e_i}\} \quad (5)$$

The relative log-polar coordinates of contour points $q_j \in \mathcal{C}$ in (1) are rewritten as follows:

$$q_j^{P_i} = (log(r_{ij}), \theta_{ij}) \quad (6)$$

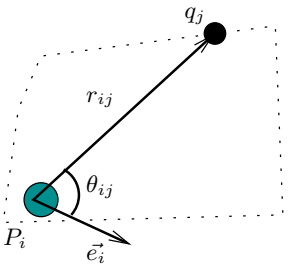

Figure 2. Relative coordinates of q_j

where P_i is the reference point, r_{ij} is the distance normalized from q_j to P_i and $\theta_{ij} =< \overrightarrow{P_i q_j}, \vec{e_i} >$ (see figure 2).

The shape context of reference point P_i is the same as defined in (2). It is a histogram of L-bins, as in [4], five bins for log(r) over the range 0.125α to 2α and 12 equally spaced radial bins (i.e. L = 60) are used. An object $\mathcal{O}$ now is described by a set of shape contexts of interest points P_i.

$$\mathcal{O} \equiv \{h_i | P_i \in \mathcal{IP}\} \tag{7}$$

3. Symbol retrieval

In [4], the distance between two shapes is measured as the symmetric sum of shape context matching costs over best matching points. This causes the complexity problem when searching the similary object from a large set of candidates. In this section, we introduce an exploitation of text retrieval technique for symbol indexing and retrieval. The objective is to reduce the complexity of on-line matchings thanks to the information pre-computed in the off-line step. The approach is similar to [16] where an efficient retrieval is achieved using a vector retrieval model including an inverted file systems based on a visual vocabulary.

3.1. Visual vocabulary construction

First, the SCIP descriptors of each symbol in the database are determined as described in section 2.2 where a descriptor is a L-vector. Next, similar descriptors are regrouped into clusters by a clustering technique. Each cluster is considered as a visual word identified by the centroid of descriptors associated and these all descriptors use the visual word as its representer. To facilitate the clustering problem, we used k-means method for current tests. The number of clusters is chosen experimentally and the distance function used is the cosinus distance.

A symbol is now described by visual words and can be treated as a text document. In figure 3 we show an example of clusters corresponding to three different visual words.

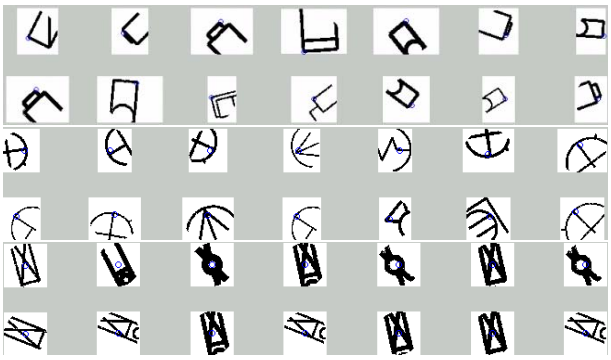

Figure 3. Example of cluster corresponding to three visual words.

3.2. Symbol retrieval using vector model

The vector model is maybe the most popular model among the research community in information retrieval. It is expected to outperform the other classic models with general collection [2]. In this model, the document is represented as a vector of word frequencies and furthermore, it is usually described by vector of weighted term frequencies whose each component provides the balance of two factors: *term-frequency (tf factor)* and *inverse document frequency term (idf factor)*. The *tf factor* shows how well that term describes the documents contents, and the *idf factor* measures the term's importance degree for distinguishing a relevant document from non-relevant one in the database.

Each SCIP descriptor of symbol *j* is matched with the nearest cluster detected in the previous section. It means that this SCIP descriptor is now considered as a visual word existing in symbol *j*, and this symbol is now considered as a document. In the next sections, the terms "document" and "symbol" refer to the same thing.

Now, we can apply the model vector to index the symbols. A symbol corresponds to a document, and a visual word corresponds to a term in model vector. Thus, a symbol *j* is represented by a *tf-idf* vector $\vec{s_j}$:

$$\vec{s_j} = \{w_{1,j}, w_{2,j}, .., w_{K,j}\} \tag{8}$$

where K is the volume of vocabulary, and $w_{i,j}$ is the weighted frequency of word *i* in document *j*:

$$w_{i,j} = tf_{i,j} * idf_i, i = \overline{1, K}$$

$$tf_{i,j} = \frac{freq_{i,j}}{max_l freq_{l,j}}$$

$$idf_i = log\frac{N}{n_i}$$

where $freq_{i,j}$ is the appearance frequence of word *i* in document *j*, N is the total number of documents in the database

and n_i is the number of documents in which the word *i* appears. $tf_{i,j}$ is defined as the normalized term frequency.

Retrieval

The *tf-ifd* vector $\vec{s}_q$ of the query symbol is computed in a similar way: computing the SCIP descriptors, then matching these descriptors with visual words and finally, determining the *tf-idf vector* corresponding to the query symbol. The degree of similarity of the query and a symbol in the database is quantified by their correlation. This correlation is measured by the cosinus distance between two vector $\vec{s}_j$ and $\vec{s}_q$.

$$sim(s_q, s_j) = \frac{\vec{s}_j \bullet \vec{s}_q}{|\vec{s}_j| \times |\vec{s}_q|} \quad (9)$$

The degrees of similarity between the query symbol and symbols in the database will be sorted in order to form a ranking list.

4. Experimental results

Since the objective of our system is the same objective of any retrieval system, i.e. to retrieve symbols in rank order with regard to the query, we use the most popular measure to evaluate the retrieval effectiveness: *precision-recall* curves [2, 7, 11, 15, 17, 20].

Recall is the fraction of the relevant symbols which have been retrieved.

$$Recall = \frac{|Ra|}{|R|}$$

Precision is the fraction of the retrieved symbols which is relevant.

$$Precision = \frac{|Ra|}{|A|}$$

where *A* is the retrieved symbols, *Ra* is the relevant symbols retrieved, $Ra \subseteq A$, and R is the relevant symbols existing in the database.

As a data set, we choose a test set of GREC[1]. It contains 300 images divided into two subsets. One is the model set which contains 50 different symbols according to 50 classes (set A), the other (set B) is a set of 250 occurences of 50 symbols classes obtained by the linear transformations (rotation et scaling) on each element of A. The occurences numbers of each class are not equal, the maximum number is 10 and the minimum is one. We choose set B (250 images) as the ground-truth set. Descriptors extracted from set B are used for building a visual vocabulary and each symbol of B will be then indexed as a document in the database. We chose experimentally K = 200 for k-means clustering, i.e. the vocabulary has 200 visual words.

[1] http://www.cvc.uab.es/grec2003/SymRecContest/

Set A plays is the set of test queries. Since the maximum number of relevant documents for each query is 10, we are only interested in the first ten documents ranked. The precision (P_r) and recall (R_r) values are calculated at each cut-off value $r = \overline{1, 10}$. We performed experiments with 50 queries in A, and took the averages of P_r and R_k for all requests at each value of *r*. The two highest curves in figure 4 are the average precision-recall curves determined from 50 queries for two values of SCIP descriptor dimension: L = 36 and 60. Some retrieval results are also shown in figure 5. As indicated in figure 4, we can obtain the results with a high precision (80%) while recall value reaches 70%. The worst precision degree is achieved (44%) when r = 10. That is, the precision and recall values are computed from the first 10 symbols retrieved but the number of correct symbols in database for each query is not always 10. So, the query for which the total number of relevant documents corresponding in the database is smaller than the number of documents retrieved, the average precision is negatively affected.

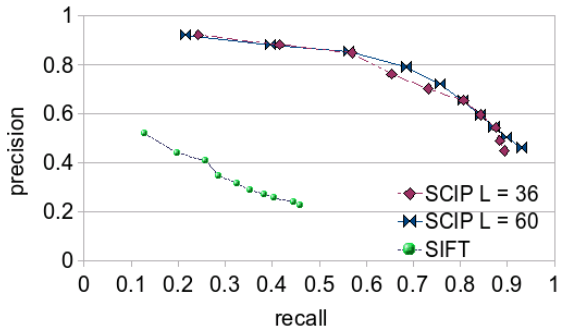

Figure 4. Retrieval effectiveness with SCIP (dimension L = 36, 60) and SIFT descriptors

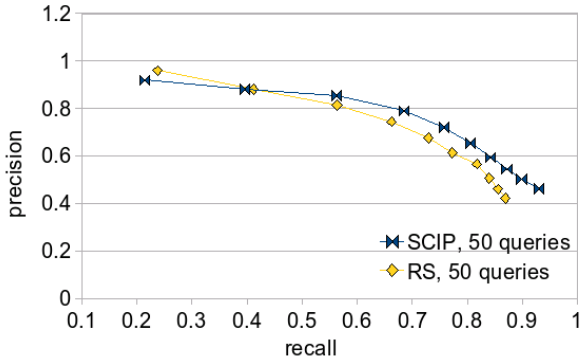

Figure 6. Retrieval effectiveness with SCIP and $\mathcal{R}$-signature

As evaluated in [11], SIFT descriptor is the best one in

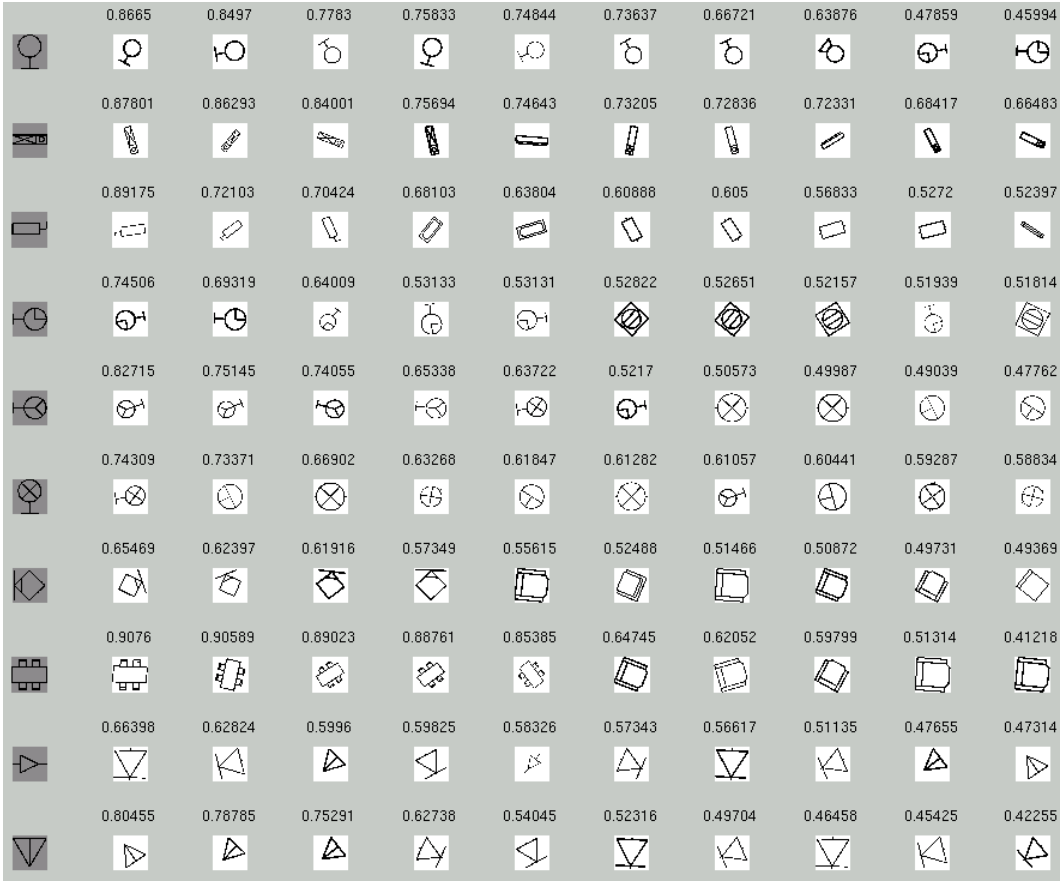

Figure 5. Retrieval exemples. The queries symbols are in the first column; other columns the nearest matches. The number of relevant symbols for each query in the database is respectively 7, 4, 3 , 5, 4, 1, 4, 5, 7, 3.

general case. Thus, we also verified the effectiveness of our descriptor when SIFT and another classical descriptor ($\mathcal{R}$-signature [19]) are used (see figures 4 and 6). The aim is not to make a comparison between two descriptors, but to show that SIFT descriptor is not adapted to graphic symbols. In addition, we can remark that the results are quite similar with the $\mathcal{R}$-signature. However, we will show below that SCIP is more robust when the goal is to retrieve incomplete symbols.

We also evaluated the system performance with other smaller values of L, i.e. smaller dimension of SCIP descriptor vector. The objective is to know how well the descriptor captures the local information and if the interest points are strong enough to describe a symbol. In this perspective, instead of using five bins for log(r) to compute the shape contexts over the range 0.125α to 2α, we chose three bins over the range 0.125α to α, the descriptor dimension is 36 (L = 36). Figure 4 shows retrieval effectiveness of the system with L = 36, 60. We found that there is no significative difference. This proves that the interest points with its descriptors can well represent the symbol.

In order to verify the adaptability of descriptors and vector model to search incomplete symbols, we tried to get the responses of our system and the $\mathcal{R}$-signature for some queries describing incomplete symbols (see figures 7 and 8).

This test indicates that the system construction strategy allows retrieval of symbols that *approximate* the query. This is an advantage for us to build in the future a retrieval system for symbols embedded into graphical documents.

5. Future works and conclusions

We have presented an adaptive solution to describe graphic symbols (SCIP descriptors) and a symbol retrieval system using the classic vector model. The SCIP descriptor is simple and invariant under rotation and scaling, it provides a good representation of local geometry at each as-

sociated keypoint. This solution is well adapted to graphic symbols. Using SCIP descriptors, we can reduce the complexity of symbol representation compared with shape contexts. In addition, building a visual vocabulary and using vector model technique allows us to reduce the complexity of matching.

The experimental results are promising but should be considered as preliminaries. When we describe a symbol by SCIP if the number of interest points is very small, the representation vector of that symbol does not guarantee a good description for the visual vocabulary construction and the matching. So, one of our future works should be a solution to add "automatically" points when the representation is too poor. Moreover, this work is the first part of a system that will not only retrieve isolated symbols but also symbols embedded into graphical documents. Future works will also be devoted to define a method for indexing and spotting symbols in a large collection of graphical documents.

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1475–1490, November 2004.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / AddisonWesley, New York, 1999.

[3] S. Belkasim, M. Shridar, and M. Ahmadi. Pattern recognition with moment invariants: a comparative study and new results. *Pattern Recognition*, 24:1117–1138, 1991.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, Avril 2002.

[5] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via plsa. In *Computer Vision ECCV 2006*, volume 3954/2006, pages 517–530. Springer Berlin / Heidelberg, May 2006.

[6] A. Jain, R. Duin, and J. Mao. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):437, 2000.

[7] F. Long, H. Zhang, and D. D. Feng. *Multimedia Information Retrieval and Management - Technological Fundamentals and Applications*, chapter Fundamentals of Contentbased Image retrieval. Spinger, 2002.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[9] S. Marshall. Review of shape coding techniques. *Image Vision Computing*, 7(4):281–294, 1989.

[10] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *International Conference on Computer Vision*, volume 1, pages 525–531, July 2001.

[11] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, October 2005.

[12] R. Prokop and A. Reeves. A survey of moment-based techniques for unoccluded object representation and recognition. *CVGIP: Graphical Models and Image Processing*, 54(5):438–460, 1992.

[13] Y. Rui, A. She, and T. Huang. A modified fourier descriptor for shape matching in mars. In *Workshop on Image Databases and Multi Media Search*, volume 8, pages 165–180, 1998.

[14] C. Schmid, R. Mohr, and C. Bauckhage. Comparing and evaluating interest points. In *In Proceedings of the ICCV, Bombay, India*, pages 230–235, 1998.

[15] L. Shapiro and G. Stockman. *Computer Vision*, chapter 8. Content-Based Image Retrieval, pages 249–273. Prentice Hall, 2001.

[16] J. Sivic and A. Zisserman. Video google: Efficient visual search of videos. In *Toward Category-Level Object Recognition*, volume 4170/2006, pages 127–144. Springer Berlin / Heidelberg, 2006.

[17] J. R. Smith. Image retrieval evaluation. In *IEEE Workshop on Content-based Access to Image and Video Databases*, page 112. IEEE Computer Society Washington, DC, USA, June 1998.

[18] S. Tabbone, L. Alonso, and D. Ziou. Behavior of the laplacian of gaussian extrema. *Journal of Mathematical Imaging and Vision*, 23(1):107–128, July 2005.

[19] S. Tabbone and L. Wendling. Recognition of symbols in grey level line drawings from an adaptation of the radon transform. In *In Proceedings of 17th International Conference on Pattern Recognition, Cambridge (UK)*, volume 2, pages 570–573, 2004.

[20] Th.Gevers and A. Smeulders. *Emerging Topics in Computer Vision*, chapter 8. Content-Based Image Retrieval: An Overview. Addison-Wesley / Prentice Hall, 2004.

[21] D. Zhang and G. Lu. Study and evaluation of different fourier methods for image retrieval. *Image and Vision Computing*, 23:3349, 2005.

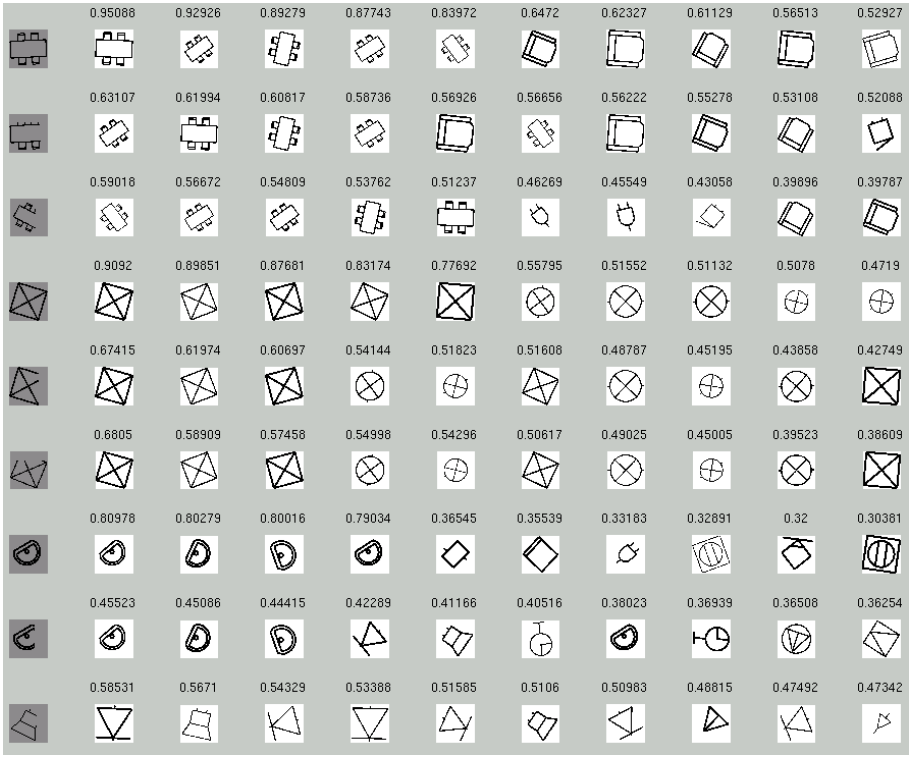

Figure 7. Results retrieved for incomplete symbols with SCIP

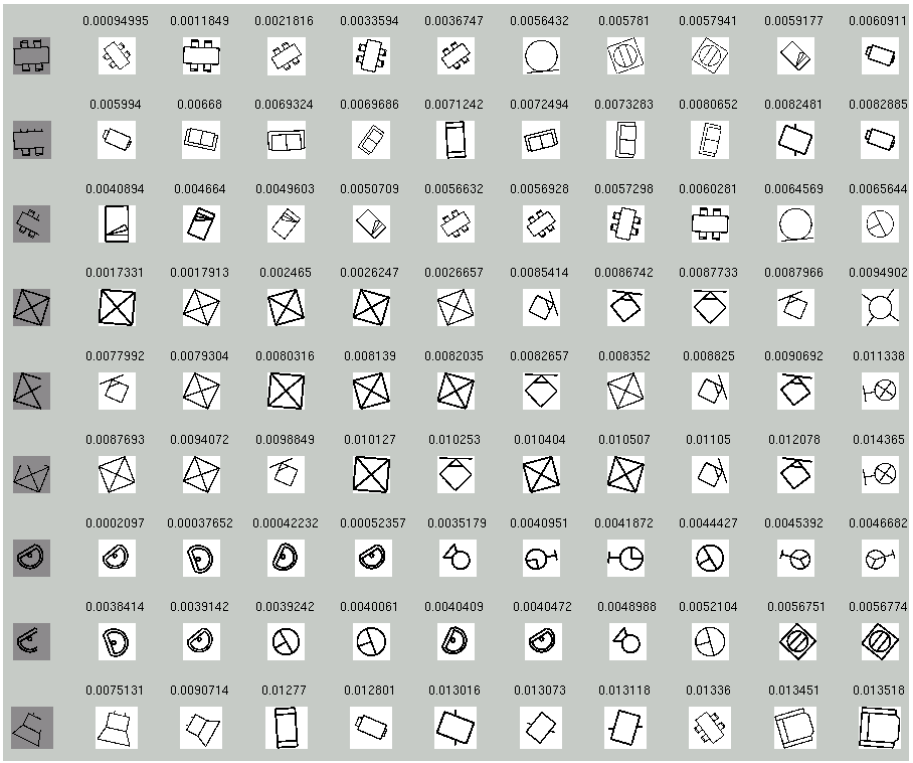

Figure 8. Results retrieved for incomplete symbols with $\mathcal{R}$-signature