Mathematical Symbol Indexing for Digital Libraries

Simone Marinai, Beatrice Miotti, and Giovanni Soda

Dipartimento di Sistemi e Informatica
University of Florence, Italy
marinai@dsi.unifi.it

Abstract. In this paper we describe our recent research for mathematical symbol indexing and its possible application in the Digital Library domain. The proposed approach represents mathematical symbols by means of Shape Contexts (SC) description. Indexed symbols are represented with a vector space-based method, but peculiar to our approach is the use of Self Organizing Maps (SOM) to perform the clustering instead of the commonly used k-means algorithm. The retrieval performance are measured on a large collection of mathematical symbols gathered from the widely used INFTY database.

1 Introduction

Nowdays, Digital library technologies are well established and understood. This is proven by the large number of papers related to this topic published in the last few years and by the broad range of systems already available in the Web. In most cases DLs deal with digitized documents, books and journals that are represented as images where Document Recognition techniques can be applied. For instance in many cases the document images are processed by means of Optical Character Recognition (OCR) techniques in order to extract their textual content.

One related research area, that has not yet fully implemented in DL systems, is based on Document Image Retrieval approaches, where relevant documents are identified relying only on image features [1].

In the last few years, most documents belonging to DLs are "born-digital" (rather than being digitized) and, as consequence, some techniques have been proposed to perform the retrieval of a user query [2] [3]. Digital Libraries can now be considered as collections of digital contents which are available through the Internet, but not necessarily public. DLs architectures and services are in continuous evolution because of their close contact with Web2.0 technologies. In the era of social networks, the users are called as main characters to build and maintain DLs which can be reached by Internet [4]. The interaction between users, to develop and update DLs, is the aim of DLs owners. Some examples of this new approach to DLs are Flickr, Facebook and other social networks which ask for the user help to increase the contents and to classify each element by means of keywords.

M. Agosti, F. Esposito, and C. Thanos (Eds.): IRCDL 2010, CCIS 91, pp. 113–124, 2010.

In our research we are interested in analyzing techniques which can be used in the phase of information extraction and retrieval from scanned or digital born documents in the Digital Library. These techniques can be classified according to the working level in three categories [5,1].

The first is the *free browsing* and is the easiest to implement: in this case a user browses through a document collection, looking for the desired information by visually checking the document images.

The second is the *recognition-based retrieval* which is based on the recognition of the document contents. According to it, the similarity between documents is evaluated at symbolic level and it assumes that a recognition engine can extract the full text of text-based documents or a set of metadata from multimedia documents. The textual information is then indexed and the retrieval is performed by means of keywords furnished by the user. The recognition-based approach has the advantage that the similarity computation and results ranking has a lower computational cost. On the other hand it has some limitations when OCR systems cannot perform well (when dealing with very noisy documents or containing multi-lingual text) or are not yet fully implemented (such as for mathematical symbol recognition, that is addressed in this paper). Some of the earliest methods adopted for the recognition-based approach, and in particular for OCR-based text retrieval, have been described in two comprehensive surveys [6,7]. A mixed approach is proposed in [8] where document image analysis techniques are used together with OCR engines and metadata extraction.

The last category is based on *recognition-free retrieval* methods and can be regarded also as content based approaches. In this case the similarity is evaluated considering the actual content of the document images. That is some features, closely related to the document images like colors texture or shape, are extracted. The user can perform the retrieval on a Query by Example (QbE) approach i.e. presenting a query image to the system and looking at a result ranking.

One advantage of a recognition-free approach is the possibility of looking for information without the need of some specific background knowledge. For example users may perform a QbE query with keywords in any language and the system does not need to know the language in the phase of the document indexing [9]. On the other hand, even recognition-free approaches have some limitations, especially regarding the selection of the feature set. Most systems work with low level features such as color, texture and shape, while only few system are able to extract high level or semantic features.

In [10], [11] and [12] key-word spotting techniques have been proposed based on the Word Shape Coding and on set of low-level features. A different approach has been proposed in [9] where words are indexed on the basis of character shapes.

Due to the large number of scientific and technical documents that are nowadays available in Digital Libraries, many efforts have been devoted to build systems which are able to recognize the mathematical expressions embedded in printed documents. Because of the very large number of symbol classes and the spatial relationships among symbol, OCR engines often fail in the recognition phase [13] [14].

According to [15], most systems for mathematical expression analysis are based on four main steps. *Layout Analysis* that is used to extract the layout of the document images. The most common techniques rely on connected components extraction and are based on bottom-up (e.g. [16]) or top-down (e.g. [17] approaches. *Symbol Segmentation* that is aimed at identifying each individual symbol, that in most cases correspond to one connected component. *Symbol Recognition* that is mostly based on machine learning techniques. For instance, Takiguchi et al. [18] and Suzuki et al. [13] represent the symbols according to pixel intensity value features and physical peculiarities. *Structural Analysis* is performed in order to understand whole equations. Toyota et al. [19] build relation trees among various symbols according to the mutual physical positions or to logic considerations.

Document image retrieval techniques have been seldom used to process mathematical expressions [20]. However, several researchers envisage the usefulness of search systems that could search for text and also for "fine-grain mathematical data" such as equations and functions [21]. Most search systems for mathematical documents rely on specific markup languages. For instance the MathWebSearch system harvests the Web for formulae indexed with MathML or OpenMath representations [22].

In this paper, we present a system based on a recognition-free approach for the retrieval of mathematical symbols belonging to a collection of documents. We do not explicitly deal with the symbol recognition, but we focus on the retrieval of mathematical symbols. This can be considered as a preliminary phase to the mathematical formulae retrieval which is the general aim of our work.

The system described in this paper is made up by three steps: in the first step for each symbol in the collection we compute a set of features that are used to index it. In the second step visual queries proposed by the user are analyzed in order to compute the same features computed during the indexing. In the last step the similarity among query vectors and each coded collection element is evaluated and the results are ranked.

The paper is organized as follows. In Sect. 2 we describe the indexing and retrieval method. The Infty database and the experiments are described in Sect. 3. Conclusions and future work are drawn in Sect. 4.

2 Mathematical Symbol Indexing

Checking two occurrences of a mathematical symbol a human observer is able to glean over the differences and basing on the symbol shape he can assert these images represent the same symbol. This kind of visual analysis should be extended to be used in a automatic process. Each image contains local interest points that concentrate most of its information, in particular the shape of the symbol is a peculiarity of the object. To demonstrate this feature of mathematical symbols we show in Fig. 1 some examples of queries with the corresponding top ten results as reported by our system.

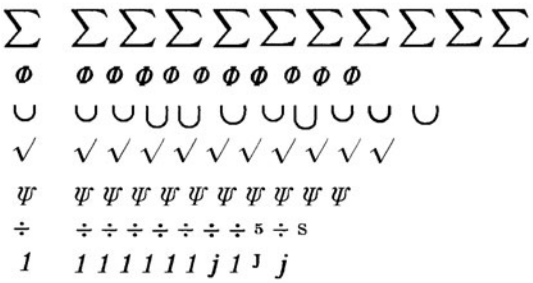


Fig. 1. Examples of queries with the first 10 retrieved symbols. The first symbol is the query.

Two main approaches can be considered to describe an image (e.g. [23]): the brightness-based ones take into account the pixel values, whereas the feature-based approaches involve the use of physical peculiarities of the symbol, such as edge elements and connected components.

In printed documents the mathematical symbols correspond in most cases to separated connected components. In our work we therefore use a feature-based approach that is more appropriate to describe the symbol shape.

Since the number of symbol classes in mathematics is very large we need a compact way to express similarity among symbols in the same class even if they look like different. The classes include characters from Greek, German, Latin alphabets and mathematical symbols. Moreover, some of them appear in general in different styles, fonts and sizes. Additional details about the dataset that we used in our experiments are reported in Sect. 3.

Among other approaches, we use Shape Contexts (SC) [23] to describe mathematical symbols shapes considering both the internal and external contours. In general, similar shapes have similar descriptors and therefore different symbols can be compared considering the SCs and then establishing a similarity measure.

The symbol image is processed to identify the internal and external contours that are subsequently described as a set of points. A subset P of sampled points is then extracted as representative of the symbol shape (Fig. 2).

2.1 Keypoints Selection

An important point of the SC-based symbol representation is the identification of keypoints on which Shape Contexts have to be evaluated. In the original paper [23] the keypoints are extracted from the internal and external symbol contours. In particular the contours are sampled with a regular spacing between keypoints. We performed the first experiments (described in Sect. 3) following this approach. We considered also other approaches to identify keypoints by looking for salient contour points. To this purpose we considered both the corner and the local maximum curvature approaches.

A corner point has two dominant and different edge directions in its neighborhood and can be detected considering the gradient of the image. In the second

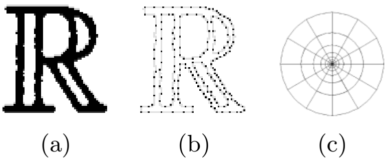


Fig. 2. (a) An example of mathematical symbol. (b) The sampled point set used to compute SCs. (c) The logarithmic mask.

approach we can define a local maximum point as a point whose local contour has a curvature which is higher than a given threshold. The curvature can be estimated considering the angle between consecutive line segments. The segments are computed with a linear interpolation of the contour on the basis of a maximum distance between the contour points and the interpolating segment. This distance has an important role in the evaluation of the curvature: with a small distance the approach will be sensible to the noise in the image and too many keypoints will be identified. On the opposite with a large distance some relevant points will be lost.

Some symbols, for example the "0", present curvature values almost constant and near to the average value while other symbols, as a "[", present only few points whose curvature values are interesting (the two corners and the endpoints). In order to alleviate the problems, the keypoints are selected among the points with a curvature greater than a threshold that is dynamically adjusted for each symbol by setting it to the average of all the curvature values in the image. In Sect. 3 we compare the results that are obtained with these approaches for keypoint selection.

2.2 Shape Contexts Evaluation

The Shape Context for each point p_i in P can be computed by considering the relative position of the other points in P. The SC for p_i is obtained by computing a coarse histogram h_i whose bins are uniform in log-polar space (Fig. 2 (b), (c)) as described in the following. Let m be the cardinality of P and p_j be one of the remaining m–1 points in P. The point p_j is assigned to one bin according to the logarithm of the Euclidean distance between p_i and p_j and to the direction of the link between p_i and p_j. The histogram h_i is defined to be the Shape Context of p_i.

The m SC vectors indirectly describe the whole symbol. It is clear that shape contexts are invariant to translations. The SC computation can be modified in order to obtain descriptions that are scale and rotation invariant [23].

For mathematical symbol indexing the rotation can be misleading because we could confuse symbols such as 6 with 9 and $\cup$ with $\cap$. To deal with mathematical symbols we also adapted the SC computation taking into account the SC radius (maximum distance of points included in the histogram) and the most set of points to be considered in the histogram population.

Large values of the SC radius allows to embrace the whole image and therefore each SC is influenced by points very far from it. With a small radius we should deal with the points that fall out of the last mask bins. Two address the latter point two alternatives are possible. In one solution all the external points are included in the last bins that will have values significantly higher than the other bins. On the opposite, if the external points are not counted the resulting SC will describe a little portion of the symbol shape with the risk to loose information.

To find out the most suitable radius we performed several experiments which are reported in detail in [24]. From these experiments it turned out that an halfway radius (20 pixels) that is a little bit smaller than the average image size should be preferred in most cases.

To increase the robustness against the symbol noise we compute h_i by counting the number of all the symbol points that belong to each bin instead of considering only the points in P. In so doing, each SC bin is more populated and more informative. We followed this approach because the symbols in the Infty dataset are small (on the average 20 x 30 pixel), and therefore the number of contour pixels is low and with the standard algorithm only a few bins of the SCs would contain some points. This choice is supported by some preliminary experiments that we described in [24].

2.3 SOM-Based Visual Dictionary

The comparison between the Shape Contexts in a query symbol and those in each indexed object can provide a very accurate evaluation of similarity among symbols. However, the computational cost of a pairwise comparison is too high and cannot be considered when dealing with large data-sets. One typical solution of this problem is based on the transformation of the shape representation using techniques adopted in the vector space model of Information Retrieval.

To this purpose, the vector quantization is first performed by clustering the vector representations and then identifying each vector with the index of the cluster it belongs to. The clusters are in most cases identified by running the K-means algorithm on a sub-set of the objects to be indexed. In the textual analogy each cluster is considered as a “visual-word” and each symbol can be represented on the basis of the frequencies of each “visual-word” in its description [25].

Although simple to implement, one limitation of the K-means clustering algorithm is that it does not take into account any similarity among clusters. In other words points belonging to different clusters (or SCs corresponding to different “visual-words”) contribute in the same way to the symbol similarity either when the clusters are similar or when they are dissimilar. The peculiarity of the approach described in this paper is the use of Self Organizing Maps (SOM) to perform the vector quantization. In this case, in contrast with K-means, the clusters are topologically ordered in the SOM map. As example we depict in Fig. 3 a portion of an SOM, used to index the mathematical symbols, together with two symbols. Each cluster is pictorially depicted by reconstructing a virtual Shape Context that corresponds to the values in the SOM related to that particular

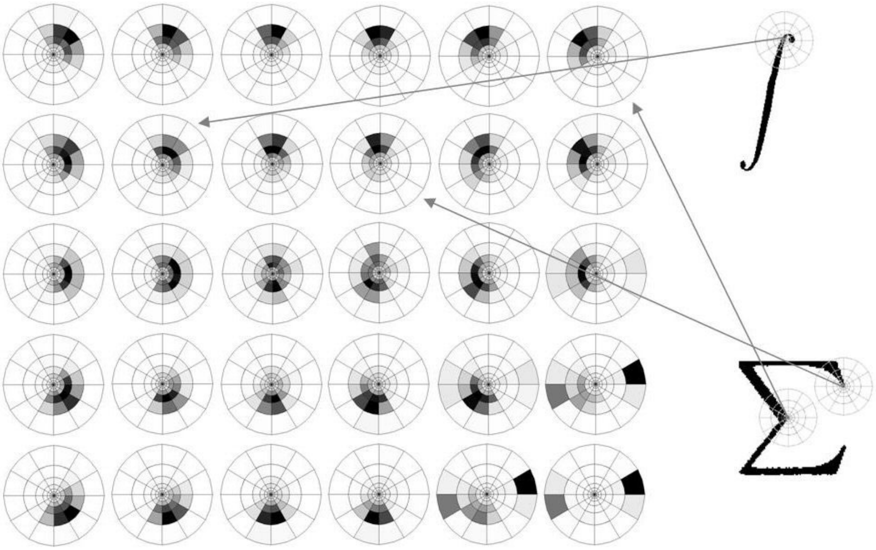


Fig. 3. Visual words obtained by SOM clustering. Each circle is a graphical representation of one cluster centroid. We show also two symbols with a reference to some visual words.

centroid. From the map it is clear that similar SCs are placed in close neurons in the map.

The similarity between the query vector and each element of the index is evaluated by means of the cosine similarity function. As proposed in [20], we have modified the cosine formula to take advantage of the topological peculiarity of the SOM map. In particular, we perform an inexact match between two vector representations of two symbols considering, for each element of the query vector that has not a correspondent in the indexed vector, the four or eight neighbors of it and we take as winner the maximum among them weighted according to its position in the map.

3 Experiments

We made our experiments on two datasets collected by the INFTY project [26]. The InftyCDB-3-A dataset consists of 188,752 symbols scanned at 400 dpi extracted from 19 articles printed by various publishers, and from other sources so as to cover all the most important mathematical symbols. The InftyCDB-3-B contains 70,637 symbols scanned at 600 dpi from 20 articles. Ground-truth information at the symbol level is provided for both datasets. It is important to notice that the same code has been assigned to different symbols that look similar (e.g. the summation symbol $\sum$ and the Greek letter Σ). In the two datasets (which consist of 346 pages) there are 393 different classes.

Table 1. Precision at 0 % Recall for the K-means and SOM experiments

Methods	K-means	SOM
SC_Std	74.69	73.17
SC_AllPoints	79.93	**86.81**

Table 2. Precision at 0 % Recall for the detection of keypoints based on local maximum curvature and corner methods. Three sizes of the SOM are compared as well.

	Curvature			Corner		
Precision at 0% Recall	10x10	10x20	20x20	10x10	10x20	20x20
sim	95.82	96.22	**97.86**	90.87	91.44	93.31
sim4	95.83	96.21	97.84	90.97	91.40	93.29
sim8	95.82	96.19	97.83	90.82	91.51	**93.44**

Before indexing the data, we computed the SC clusters on a set of 22, 923 symbols, belonging to 53 pages randomly selected from the whole dataset. From each symbol we extract around 50 SCs so that we used 1,102,049 feature vectors for clustering. We then indexed all the 259,357 symbols in the two datasets and we performed several experiments to compare alternative approaches that can be used to index the data.

To evaluate the retrieval results we use the *Precision-Recall* curves and a single numerical value: the Precision at 0 % Recall, which is obtained through an interpolation procedure of the Precision-Recall curve as detailed in [27]. As usual, the Recall is defined as the fraction of the relevant symbols which have been retrieved: $Recall = \frac{|tp|}{|tp+fn|}$; the Precision is defined as the fraction of retrieved symbols which are relevant: $Precision = \frac{|tp|}{|tp+fp|}$; where *tp* (true positive) are the retrieved symbols which are relevant, *fp* (false positive) are the retrieved symbols which are not relevant, *fn* are the relevant symbols which have not been retrieved. We computed the P-R curve for interpolation after estimating the precision when the recall is a multiple of 10.

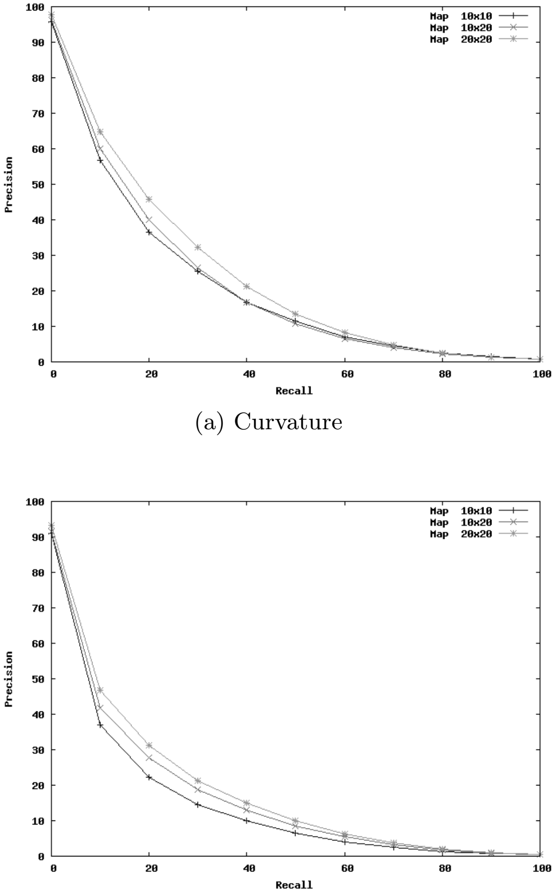
In the experiments reported in this paper we made 392 queries randomly selected from the dataset. To obtain a correct comparison among methods, we always used the same set of 392 queries. To estimate the suitability of the SOM clustering we first compared it with the K-means clustering. Some preliminary experiments are reported in [20] and in [24]. The latter are summarized in Table 1 where we can verify that the SOM clustering together with a computation of SCs with all the symbol points (SC_AllPoints) provides the best results. To compare

Table 3. Area under the curve (AUC) of P-R in case of 3920 queries

	sim	sim4	sim8
AUC	2345,08	**2410,34**	2383,76

the two approaches for keypoint selection described in Sect. 2.1 we performed some experiments where we used the same settings of the previous experiments.

In Table 2 and Fig. 4 we show the Precision at 0 % Recall and the Precision-Recall curves for the two approaches. In the experiments we considered three map sizes (with 100, 200, and 400 neurons) and also three functions to compute



(a) Curvature

(b) Corner

Fig. 4. Precision-Recall curves related to Table 2

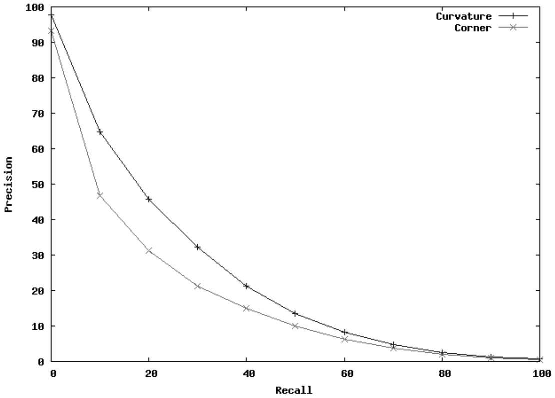


Fig. 5. Comparison of Precision-Recall curves

the symbol similarity. The similarity among query vectors and indexed elements is evaluated by means of the cosine similarity function and two variants of it *sim*4 and *sim*8 as explained in detail in [20]. From these experiments we can observe that larger maps provide in general better results. However, in this experiment there is no considerable difference among the three similarity functions. To compare the two methods we report in Fig. 5 the best plots of Fig. 4. From this figure it is clear that the curvature method is better than the other.

From the previous experiments it is not possible to understand whether there is any advantage in the use of one similarity function with respect to the others. This is probably due to the low number of queries that we used. We therefore performed an additional experiment considering an SOM map with 400 centroids and the curvature approach. The other settings are fixed as before, but we performed 3920 queries. The results, evaluated with the Precision Recall curve, show that the values of Precision at 0 % Recall are nearly the same for all similarity functions. To compare in a better way the methods, we considered the area under these curves as a quality measure. The results are shown in Table 3. As we can see, the *sim*4 similarity function although starting from a similar value than the others, tends to have higher values when the Recall increases.

4 Conclusions

In this paper, we proposed a technique for image-based symbol retrieval based on Shape Context representation encoded with a *bag of visual words* method. The peculiarity of the approach is the use of Self Organizing Maps for clustering the Shape Contexts into a suitable visual dictionary. We also compared various methods to compute the SCs as well as different clustering approaches. Experiments performed on a large and widely used data set containing both alphanumeric and mathematical symbols allows us to positively evaluate the proposed approach.

Future work includes a deeper comparison of various retrieval approaches. We aim also to extend the retrieval mechanism to incorporate the structural information of the formulae in the retrieval algorithm.

References

1. Marinai, S.: A Survey of Document Image Retrieval in Digital Libraries. In: Sulem, L.L. (ed.) Actes du 9ème Colloque International Francophone sur l'Ecrit et le Document, SDN 2006, pp. 193–198 (September 2006)
2. Chen, N., Shatkay, H., Blostein, D.: Use of figures in literature mining for biomedical digital libraries. In: Proc. DIAL, pp. 180–197 (2006)
3. Esposito, F., Ferilli, S., Basile, T., Mauro, N.D.: Automatic content-based indexing of digital documents through intelligent processing techniques. In: Proc. DIAL, pp. 204–219 (2006)
4. Gazan, R.: Social annotations in digital library collections. DLib. Magazine 14(11/12) (2008)
5. Wan, G., Liu, Z.: Content-based information retrieval and digital libraries. Information Technology & Libraries 27, 41–47 (2008)
6. Doermann, D.: The indexing and retrieval of document images: A survey. Computer Vision and Image Understanding 70(3), 287–298 (1998)
7. Mitra, M., Chaudhuri, B.: Information retrieval from documents: A survey. Information Retrieval 2(2/3), 141–163 (2000)
8. Belaïd, A., Turcan, I., Pierrel, J.M., Belaïd, Y., Hadjamar, Y., Hadjamar, H.: Automatic indexing and reformulation of ancient dictionaries. In: DIAL 2004: Proceedings of the First International Workshop on Document Image Analysis for Libraries, Washington, DC, USA, p. 342. IEEE Computer Society, Los Alamitos (2004)
9. Marinai, S., Marino, E., Soda, G.: Font adaptive word indexing of modern printed documents. IEEE Transactions on PAMI 28(8), 1187–1199 (2006)
10. Bai, S., Li, L., Tan, C.: Keyword spotting in document images through word shape coding. In: ICDAR 2009: Proceedings of the Tenth International Conference on Document Analysis and Recognition, p. 331. IEEE Computer Society, Los Alamitos (2009)
11. Li, L., Lu, S.J., Tan, C.L.: A fast keyword-spotting technique. In: ICDAR 2007: Proceedings of the Ninth International Conference on Document Analysis and Recognition, Washington, DC, USA, pp. 68–72. IEEE Computer Society, Los Alamitos (2007)
12. Lu, S., Li, L., Tan, C.L.: Document image retrieval through word shape coding. IEEE Trans. Pattern Anal. Mach. Intell. 30(11), 1913–1918 (2008)
13. Suzuki, M., Tamari, F., Fukuda, R., Uchida, S., Kanahori, T.: Infty: an integrated ocr system for mathematical documents. In: DocEng 2003: Proceedings of the 2003 ACM Symposium on Document Engineering, pp. 95–104. ACM, New York (2003)
14. Garain, U., Chaudhuri, B.B., Chaudhuri, A.R.: Identification of embedded mathematical expressions in scanned documents. In: ICPR, vol. 1, pp. 384–387 (2004)
15. Guo, Y., Huang, L., Liu, C., Jiang, X.: An automatic mathematical expression understanding system. In: ICDAR 2007: Proceedings of the Ninth International Conference on Document Analysis and Recognition, Washington, DC, USA, vol. 2, pp. 719–723. IEEE Computer Society, Los Alamitos (2007)

16. Anil, K.J., Bin, Y.: Document representation and its application to page decomposition. IEEE Trans. Pattern Anal. Mach. Intell. 20(3), 294–308 (1998)
17. Chang, T.Y., Takiguchi, Y., Okada, M.: Physical structure segmentation with projection profile for mathematic formulae and graphics in academic paper images. In: ICDAR 2007: Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Washington, DC, USA, vol. 2, pp. 1193–1197. IEEE Computer Society, Los Alamitos (2007)
18. Takiguchi, Y., Okada, M., Miyake, Y.: A study on character recognition error correction at higher level recognition step for mathematical formulae understanding. In: 18th International Conference on Pattern Recognition, ICPR 2006, vol. 2, pp. 966–969 (2006)
19. Toyota, S., Uchida, S., Suzuki, M.: Structural analysis of mathematical formulae with verification based on formula description grammar. In: Bunke, H., Spitz, A.L. (eds.) DAS 2006. LNCS, vol. 3872, pp. 153–163. Springer, Heidelberg (2006)
20. Marinai, S., Miotti, B., Soda, G.: Mathematical symbol indexing using topologically ordered clusters of shape contexts. In: Int'l. Conference on Document Analysis and Recognition, pp. 1041–1045 (2009)
21. Youssef, A.: Roles of math search in mathematics. In: Borwein, J.M., Farmer, W.M. (eds.) MKM 2006. LNCS (LNAI), vol. 4108, pp. 2–16. Springer, Heidelberg (2006)
22. Kohlhase, M., Sucan, I.: A search engine for mathematical formulae. In: Calmet, J., Ida, T., Wang, D. (eds.) AISC 2006. LNCS (LNAI), vol. 4120, pp. 241–253. Springer, Heidelberg (2006)
23. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(4), 509–522 (2002)
24. Marinai, S., Miotti, B., Soda, G.: Mathematical symbol indexing. In: AI*IA 2009: Proceedings of the XIth International Conference of the Italian Association for Artificial Intelligence Reggio Emilia on Emergent Perspectives in Artificial Intelligence, pp. 102–111. Springer, Heidelberg (2009)
25. Yang, J., Jiang, Y.G., Hauptmann, A.G., Ngo, C.W.: Evaluating bag-of-visual-words representations in scene classification. In: MIR 2007: Proceedings of the international workshop on multimedia information retrieval, pp. 197–206. ACM, New York (2007)
26. Suzuki, M., Tamari, F., Fukuda, R., Uchida, S., Kanahori, T.: Infty: an integrated ocr system for mathematical documents. In: DocEng 2003: Proceedings of the 2003 ACM symposium on Document engineering, pp. 95–104. ACM, New York (2003)
27. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley, Reading (1999)