Plane Geometry Figure Retrieval with Bag of Shapes

Conference Paper · April 2014

DOI: 10.13140/2.1.3802.4325

6 authors, including:

lu Liu
National University of Singapore

Xiaoqing Lyu
Peking University

Jingwei Qu
Peking University

Liangcai Gao
Peking University

2014 11th IAPR International Workshop on Document Analysis Systems

Plane Geometry Figure Retrieval with Bag of Shapes

Lu Liu, Xiaoqing Lu, Keqiang Li, Jingwei Qu, Liangcai Gao, Zhi Tang
Institute of Computer Science & Technology, Peking University, Beijing, China
State Key Laboratory of Digital Publishing Technology, Beijing, China
{liulu.pku, lvxiaoqing, chocolate, qujingwei, gaoliangcai, tangzhi}@pku.edu.cn

***Abstract*—Digital education is serving an increasingly important function in most educational institutions, thus resulting in the production of a large number of digital documents online for education purposes. However, convenient ways to retrieve mathematic geometry questions are lacking because current retrieval systems largely rely on keywords instead of geometry figure images. This study focuses on plane geometry figure (PGF) image retrieval with the aim of retrieving relevant geometry images that contain more structural information than a question text stem. To fully use geometrical properties, a Bag-of-shapes (BoS) method is proposed to build the feature descriptor of an image. The BoS method contains either basic geometric primitives or dual-primitive structures along with several specific geometrical features for shape description. Based on the BoS feature descriptor, we apply cosine similarity with group feature weight as vector similarity measure for ranking to achieve high efficiency. For a PGF image query, the retrieval results are provided in an appropriate ranking order, which has high visual similarity with respect to human perception. Retrieval experiments and evaluation results show the effectiveness and efficiency of the proposed BoS shape descriptor.**

Keywords: plane geometry figure, bag-of-shapes, geometric structure, shape feature, image retrieval

I. INTRODUCTION

A growing number of teaching materials have recently been digitalized. These materials can be accessed through the Internet. A large number of digital questions and examination papers exist online, particularly for mathematical education. However, people remain confused when searching for a specific geometry question. To our knowledge, existing plane geometry question retrieval systems are generally based on keywords in question stems. However, the description of a few keywords can hardly cover the main points of mathematical questions, let alone retrieve similar questions. As a result, retrieval precision and recall might be unsatisfactory.

Motivated by the lack of a vertical search engine for plane geometry figures, a plane geometry figure (PGF) image (Figure 1) retrieval system for geometry learning and teaching is proposed. In content-based image retrieval (CBIR), color, texture, and shape are important visual features for describing an image. However, the greatest challenge of such PGF images is that they are only composed of lines without color, texture feature. Furthermore, the basic geometric primitives come with different relationships and complicated layouts, such as overlapping, embodying, and sharing common lines or angles. Traditional image recognition methods often fail to describe and analyze such PGF images because of the lack of low-level local features. It is required to focus on the inter-structure and primitives' layout instead of the fixed pattern compared to domain-symbol retrieval. The bag-of-words (BoW) and bag-of-features methods are widely used for image retrieval system. Based on these two methods, we propose the bag-of-shapes (BoS) method to detect and build shape feature descriptors automatically for plane geometry image retrieval.

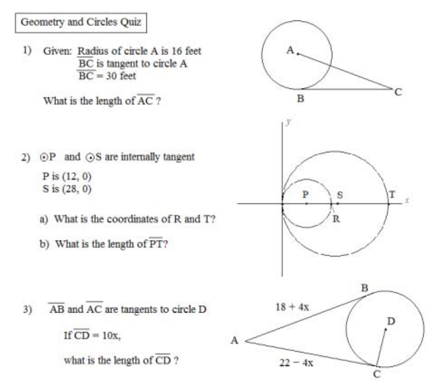

Fig. 1. Examples of plane geometry questions with PGF images

Shape-based image retrieval often involves shape descriptors based on skeleton, boundary, local region, or whole image. Shape description techniques can be broadly categorized into two groups: contour-based and region-based shape descriptors [1, 2], which represent shapes in an image either based on interior region or boundary. Several sketch retrieval systems such as MindFinder [3] are similar to PGF retrieval due to the line-drawn query images, but most sketch-based retrieval methods rely on the writing sequence of strokes other than shape features. Many of the shape-based techniques are applied in traffic sign detection [4,5,6], architectural symbol spotting [7], electronic device symbol retrieval [8,9] and trade-mark logo retrieval [10,11].

Bob Futrelle has done much research on document diagram recognition and understanding [12,13]. They extract graphics primitives using PDF metadata and define several *graphemes* (vertical-tick, horizontal-tick, line segment, curve, callout, adjacent rectangles, bars, branch, data point) which are made up of two primitives with geometrical constraints. Finally, classify the diagram's category. Similar to the work, we aim to retrieve PGF images, detect geometrical shapes from raster images, re-define the dual-primitive spatial structure (*graphemes*) and calculate the similarity between PGF images.

978-1-4799-3243-6/14 $31.00 © 2014 IEEE
DOI 10.1109/DAS.2014.53

This paper is organized as follows: Section II provides an outline of the retrieval system. Section III elaborates the feature extraction procedure. Section IV describes the retrieval step utilizing cosine similarity for Vector Similarity Measure. Section V presents the experiment and evaluation results. Section VI discusses the conclusion and future work.

II. BoS RETRIEVAL OUTLINE

In order to build PGF image database, we collect a number of mathematical digital documents (e.g., PDF), which contain question subjects and PGF images, as well as PGF images online. Then we use the method [14] to extract most PGF images automatically for the establishment of our image database. We also crawled a large set of PGF images online for further experiment.

The overall workflow is shown in Figure 2. First, to extract the semantic-level shape feature, we need to decompose the image into several basic geometric primitives, including circle, triangle, rectangle, polygon, parallelogram, and trapezoid, using a series of recognition and detection techniques. Second, we analyze the structural relationship among the primitives to identify a strong-relevant structure. Third, based on BoS and structures, we conduct feature engineering to build the feature descriptor. Fourth, we apply cosine similarity with group feature weight for ranking in retrieval.

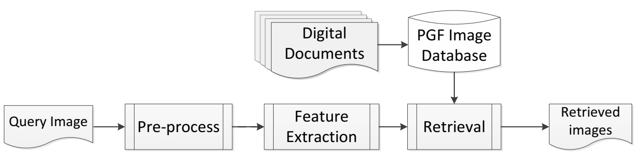

Fig. 2. Workflow of PGF retrieval

A feature vector (*FV*) is the BoS feature descriptor of an image, composed of five types of features: single geometric primitive histogram, dual-primitive structure binary feature, main primitive feature, global feature, and boundary curvature (scaled), as in Eq.(2).

$$\mathrm{FV} = [\mathrm{FV_{primitive}}, \mathrm{FV_{structure}}, \mathrm{FV_{main}}, \mathrm{FV_{global}}, \mathrm{FV_{boundary}}] \quad (2)$$

III. PGR FEATURE EXTRACTION

A. Preprocess

For an input image, several pre-process methods are adopted. After removing the character labels of points, which are small connected components, we obtain a pure shape of the PGF image. The image is converted into a binary black/white image and then dilated and eroded to fill gaps. Finally, a thinning procedure is adopted to obtain the skeleton image.

B. Basic Geometric Primitive Feature

A PGF image *I* is composed of several basic geometric primitives, which are considered as line segments, arcs, triangles, circles, rectangles, parallelogram, and trapezoid, denoted as the BoS (terms) $\{S_1, S_2, \ldots, S_k\}$. Given a PGF, the number of each shape, S_i, has to be determined.

Basic lines are detected through Hough transform to identify line segments. Circles are detected using the randomized circle detection method [15]. Rectangle, parallelogram, and trapezoid are detected using the proposed symmetry analysis algorithm [16]. Triangles are detected using key-points and edge tracking based on the geometry restrictions of a triangle. Finally, unused line segments (straight lines and arcs) are added to the corresponding S_i. We consider 10 kinds of shapes: line, triangle, trapezoid, rectangle, regular polygon (5-8 sides), parallelogram and circle, result in a 10-dimension single primitive binary feature vector.

Feature vector $FV_{primitive}$ is built in TF-IDF (term frequency–inverse document frequency) form, which is often used as a weighting factor in information retrieval and text mining. For instance, PGF is a document, each shape feature S_j is a term, N is the total number of *I*, for an image I_i, $n(i,j)$ is the number of term S_j in I_i, and $df(j)$ is the number of images that contain term S_j. Then, $FV_{primitive}$ for image i and term j is calculated as

$$FV_{primitive}(i,j) = tf(i,j) * idf(j) = \frac{n(i,j)}{\sum_k n(i,k)} * \log\left(\frac{N}{df(j)}\right) \quad (1)$$

C. Dual-primitive Structure

Considering that basic single primitive sets alone are insufficient to describe the layout of PGF, particularly the spatial relationship among all primitives, we introduce several compound shapes based on the strong spatial correlation. The relationship of the three main shape types: triangle, circle, and rectangle, may be overlapping or tangent (inside, outside), including sharing end points. From the experiment, 41 strong-correlated compound shapes are selected as the dual-primitive structures, as shown in Figure 3. $FV_{structure}$ is built as frequency histogram (scaled to 0~1), with 41 dual-primitive terms. More details can be found in our paper [17].

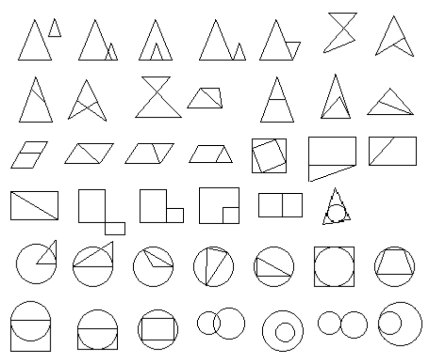

Fig. 3. The 41 dual-primitive terms

D. Main Primitive detailed feature

Apparently, the dominant primitive is considered to be more salient than small ones. To emphasize the importance of top k (k=3) main primitives, we introduce the shape attributes of the top 3 largest (in area) primitives. Four categories of features are calculated: shape type, shape area and circumstance, and shape attributes.

$$FV_{main} = [FV_{type}, FV_{area}, FV_{circum}, FV_{attributes}] \quad (3)$$

FV_{type} is a 10 dimensional binary feature vector, there is only one dimension equals to 1, which is the corresponding type dimension and others are 0. We also consider 10 kinds of shapes as above mentioned. FV_{area} and FV_{circum} are the area and circumstance value scaled by the bounding box's area/circum. $FV_{attributes}$ includes four cases as follows: ①If the type is triangle, we obtain six features: equilateral triangle, isosceles triangle, vertical triangle, arbitrary triangle, max angle proportion, and max edge proportion. ②If the type is circle, we obtain two features: half-circle and diameter length proportion. ③If the type is rectangle, we detect whether the shape is a square. ④If the type is trapezoid, we determine whether the shape is isosceles, right-angled, or arbitrary. The max edge length and max angle proportion are also considered.

Finally, we obtain a 26-dimension FV_{main} feature for each of the top 3 main shapes, then 78 dimensions overall.

E. Global Feature

Some of the classical shape description metrics are also considered as global feature to improve feature robustness. Eight kinds of global metrics are calculated to build the global feature.

1) Area ratio average of all graphics primitives

$$FV_{global\_1}(i) = \frac{\sum_{j=1}^{n_i} A_j{}^i}{A_{BB}^i * n_i} \quad (4)$$

where $A_j{}^i$ is the area of primitive S_j in image I_i, $j = 1,2,\dots,n_i$, n_i is the number of primitives in image I_i, and A_{BB}^i is the area of the bounding box of I_i.

2) Circumference ratio average

$$FV_{global\_2}(i) = \frac{\sum_{j=1}^{n_i} C_j{}^i}{C_{BB}^i * n_i} \quad (5)$$

where $C_j{}^i$ is the circumference of primitive S_j in image I_i, $j = 1,2,\dots,n_i$, n_i is the number of primitives in image I_i, and C_{BB}^i is the circumference of the bounding box of I_i.

3) Centroid distance average

TABLE I. CENTROID DEFINITION

Primitive type	Centroid
Circle	Center of circle
Triangle	Inner center
Rectangle	Intersection point of diagonal
Polygon	Mean coordinates

$$FV_{global\_3}(i) = \frac{\sum_{k=1}^{n_i}\sum_{j=1}^{n_i} D_{k,j}}{COM_{n_i}^2} \quad (6)$$

$D_{k,j}$ is the distance between two centroids of primitives S_k and S_j in image I_i, $k, j = 1,2,\dots,n_i, k \neq j$, n_i is the number of primitives in image I_i, and $COM_{n_i}^2$ is the combination number.

4) Centroid distance average inner type

$$FV_{global\_4}^t(i) = \frac{\sum_{k=1}^{n_{i_t}}\sum_{j=1}^{n_{i_t}} D_{k,j}}{COM_{n_{i_t}}^2} \quad (7)$$

$$t \in \{triangle, circle, other\}$$

The centroid distance inner type is calculated considering triangle, circle, and others (mainly rectangle).

5) Circularity

$$FV_{global\_5}(i) = \frac{4\pi}{n_i} * \sum_{j=1}^{n_i} \frac{A_j{}^i}{(C_j{}^i)^2} \quad (8)$$

where $A_j{}^i$ is the area of primitive S_j in image I_i, $C_j{}^i$ is the circumference of primitive S_j, $j = 1,2,\dots,n_i$, n_i is the number of primitives in image I_i.

6) Compactness

$$FV_{global\_6}(i) = \frac{1}{4\pi n_i} * \sum_{j=1}^{n_i} \frac{(C_j{}^i)^2}{A_j{}^i} \quad (9)$$

where $A_j{}^i$, $C_j{}^i$, and n_i are the as above.

7) Minimum/maximum axis length average

TABLE II. MINIMUM/MAXIMUM AXIS DEFINITION

Primitive type	Minimum/Maximum axis
Circle	Diameter
Triangle	Angular bisector (min/max)
Polygon	Diagonal (min/max)

$$FV_{global\_7}(i) = \frac{\sum_{j=1}^{n_i} L_j{}^i}{n_i} \quad (10)$$

where $L_j{}^i$ is the length of the minimum/maximum axis.

8) Normalized radial length

$$FV_{global\_8}(i) = [\mu_i, \sigma_i] \quad (11)$$

$$\mu_i = \frac{1}{n_{B_i}} * \sum_{j=1}^{n_{B_i}} D_{B_i}$$

$$\sigma_i = \sqrt{\frac{1}{n_{B_i}} * \sum_{j=1}^{n_{B_i}} (D_{B_i} - \mu_i)}$$

B_i is the point in the boundary of I_i, and n_{B_i} is the number of the boundary points. D_{B_i} is the distance between point B_i and the centroid point of I_i. Then, we derive the mean μ_i and standard deviation σ_i of the normalized radial length of I_i.

F. Boundary Feature

The boundary is an outstanding feature that can never be overlooked, particularly when detection failures or mistakes might be present. Although the boundary feature is insufficient to describe the inner structure, it is still a significant feature for the robustness of the algorithm due to the detection errors or basic geometry shape missing situations.

We use the envelope extraction method [18] to extract the boundary of the PGF image and convert the boundary into a curvature description. All curvatures are then aligned from the curvature peak index bin and then scaled into the longest curvature length. In this way, the boundary feature is rotational-invariant, translation-invariant, and scale-invariant without changing the curvature distribution.

To emphasize the sharp points and bumps, we introduced four other curvature metrics: mean, standard deviation, and peak number. We also use K-means clustering algorithm to cluster the boundary features to bring in a priori knowledge. The cluster number is then recorded as a binary feature, which indicates that the corresponding cluster feature value is one; otherwise, zero. In this way, the boundary feature of I_i is expressed as

$$\mathrm{FV}_{\mathrm{boundary}}(i) = [curvature_i, mean_i, std_i, peakNum_i, clusterNo_i] \quad (12)$$

Based on all the above five types of features, the feature descriptor of each PGF image - FV as in Eq.(2) has been established. Last step is normalization using max/min method.

IV. PGR RETRIEVAL

Given a query image, the feature extraction step builds the feature for the query. Another important problem must be carefully defined that is what is the relevancy for a PGF image? Is this relevancy merely a true or false definition or a ranking problem? According to human visual perception, such as the Gestalt Theory [19], proximity, similarity, closure, symmetry, continuation, smallness, surroundedness, and simplicity imply the structure and organization of wholes. Theoretically, as regards the geometry problem, we consider six main factors of geometric primitives: (1) type, (2) number, (3) shape similarity, (4) dominancy, (5) topology structure, and (6) boundary. Each of the factors works in various circumstances at various degrees. Considering the above relevancy factors, we conduct feature engineering in the ranking model.

Considering high efficiency and low latency requirement, we choose Cosine similarity to measure the relevance between a query image and candidate. The performance of each group of feature is shown in Figure 4, measured by precision average at top k. It shows that single primitive feature is outstanding, global and boundary feature keep a moderate level, and structure and main primitive feature might be unsatisfactory, because the number of dual-primitive structure in each image is limited, and there may be no outstanding main primitive.

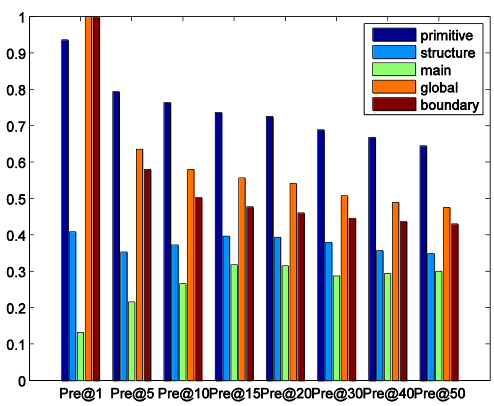

Fig. 4. The precision at top k of the five groups of feature

V. EXPERIMENT AND EVALUATION

The experiment environment included an Intel Core i5 (3.20 GHz) processor with 4.00G RAM and Windows 7 operating system as well as Matlab 2011b. Two PGF image databases were obtained. Database 1 contains 267 black/white PGF images extracted from digital pdf documents, size range from 60*96 to 400*96. Database 2 contains 1519 images collected from math-learning websites. As we had collected the ranking ground-truth labeled for database 1, so we used database 1 to conduct evaluation experiments. We also established an online website named GFfinder (Geometry Figure Finder) http://59.108.48.30/gffinder.htm, based on our PGF retrieval system.

A. Retrieval

The top 23 retrieval results of 2 queries (denoted with red square) in database 2 are shown in Figure 5 and 6. From experiment, we see that our algorithm works well in PGF image, especially with two intersected circles, circles overlapped with triangles, or several overlapped triangles. The results show that our descriptor better describes the inner structures of the PGF images.

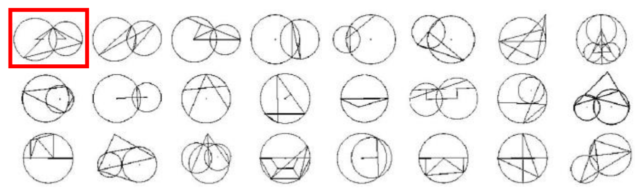

Fig. 5. Top 23 retrieval results for query 1 (left to right, top to down)

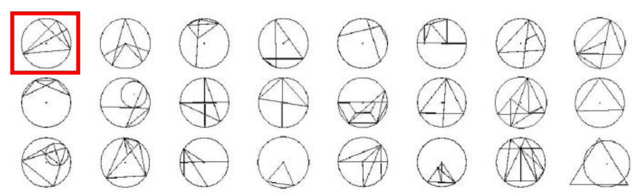

Fig. 6. Top 23 retrieval results for query 2 (left to right, top to down)

B. Evaluation

We use three classical evaluation methods to evaluate the retrieval quality: precision, mean average precision (MAP), and normalized discount cumulative gain (NDCG) [20], as shown in Figure 7. The ground-truth of database 1 is conducted through an online questionnaire. In order to evaluate the ranking performance, it is required to label the similarity degree between every two images with 0~5 score. Higher score stands for higher similarity and 0 means not relevant at all. We have labeled the similarity matrix for the 267 database, which containing C_{267}^2 number of similarity value. We also manually revise the labeled answers in a list-wise manner to ensure that the labels are accurate and reliable.

MAP consumes two categories: relevant/irrelevant. Relevant documents should be ranked before irrelevant ones. Thus, we consider a labeled similarity score greater than 0 to be relevant and others as irrelevant. Given a ranking list of query $\mathrm{q_i}$, average precision is defined as the average of precision after each positive instance is retrieved, which is calculated as

$$\mathrm{P(i)} = \frac{\sum_{\pi(k)\le\pi(i)} pos(q,k)}{\pi(i)} \quad (13)$$

$$\mathrm{AP(q)} = \frac{\sum_{i=1}^{n_q} P(i)*pos(q,i)}{\sum_{i=1}^{n_q} pos(q,i)} \qquad \mathrm{MAP} = \frac{\sum_q AP(q)}{\#q}$$

Where $\pi(i)$ is the rank position of $\mathrm{I_i}$, and $pos(q,i)$ is a label indicator of whether k is relevant to q.

NDCG measures the multi-degree in relevance ranking, wherein highly relevant documents should be ranked before low relevant ones, which is calculated as

$$\mathrm{DCG} = \sum_{i=1}^{n} \frac{2^{pos(q,i)}-1}{\log(\pi(i)+1)} \quad \mathrm{NDCG} = \frac{DCG}{DCG_{max}} \qquad (14)$$

Where $\mathrm{DCG_{max}}$ is the maximum value of the ideal ordering.

We give the precision, MAP and NDCG curve of database 1 against recall and rank respectively as shown in Figure 7.

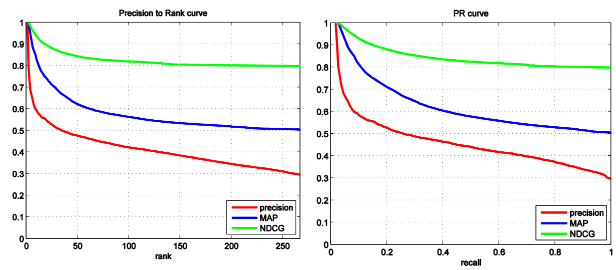

Fig. 7. (a) Precision, MAP and NDCG curve (b) The precision and recall curve

VI. CONCLUSION

In this study, we focuse on PGF image retrieval with the aim of retrieval by mathematic images that contain more structure information than several keywords in the question text stem. A BoS feature descriptor is proposed that contains both basic geometric primitives and dual-primitive structures, along with some specific geometrical property and boundary curvature feature. Yet, in this paper, we don't show the algorithm comparison result due to the specific problem and experiment database. In the future, we will introduce machine learning method and feature engineering techniques to further improve the PGF image retrieval function and performance.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No.61202232) and National Key Technology R&D Program of China (No.2012BAH40F01).

REFERENCES

[1] D. Zhang, and G. Lu, "Review of shape representation and description techniques," Pattern Recognition, vol. 37, no. 1, pp.1-19, 2004.

[2] Ryszard S. Choras, "Shape-based image retrieval," in Proceedings of the 7th WSEAS International Conference on Signal Processing (SIP),pp.99-104,2008.

[3] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, L. Zhang, "MindFinder: interactive sketch-based image search on millions of images," in Proceedings of the international conference on Multimedia, pp. 1605-1608,2010.

[4] J. He, and Y. Ma, "Triangle detection based on windowed Hough Transform," in 7th International Conference on Wavelet Analysis and Pattern Recognition, pp. 95 -100, 2009.

[5] M. Garcia-Garrido, M. Sotelo and E. Martin-Gorostiza, "Fast road sign detection using Hough Transform for assisted driving of road vehicles," in 10th International Conference on Computer Aided Systems Theory, pp. 543-548, 2005.

[6] W. Kuo, and C. Lin, "Two-Stage road sign detection and recognition," in IEEE International Conference on Multimedia and Expo, pp.1427-1430, 2007.

[7] S. Ahmed, M. Liwicki et. al., "Automatic Room Detection and Room Labeling from Architectural Floor Plans," in 10th IAPR International Workshop on Document Analysis Systems (DAS),2012.

[8] N. Nayef and T. M. Breuel, "Graphical symbol retrieval using a branch and bound algorithm," in Proceedings of 2010 IEEE 17th International Conference on Image Processing, pp. 2153-2156, September 2010.

[9] T. O. Nguyen, S. Tabbone, O. R. Terrades, "Symbol descriptor based on shape context and vector model of information retrieval," in The Eighth IAPR Workshop on Document Analysis Systems (DAS),2008.

[10] Jain, Rajiv, and David Doermann. "Logo retrieval in document images," Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on. IEEE, 2012.

[11] Rusiñol, M., & Lladós, J. "Efficient logo retrieval through hashing shape context descriptors," Proceedings of the 9th IAPR International Workshop on Document Analysis Systems (DAS). ACM, 2010.

[12] Shao, M. & Futrelle, R. P. "Recognition and Classification of Figures in PDF Documents," In W. Liu and J. Lladós (Eds.): Selected papers from GREC 2005, LNCS 3926, pp. 231-242, 2006.

[13] Futrelle, R. P., Shao, M., et. al. "Extraction, layout analysis and classification of diagrams in PDF documents," In ICDAR 2003 (Intl. Conf. Document Analysis & Recognition), pp. 1007-1014, Scotland, 2003.

[14] C. Xu, Z. Tang, X. Tao, and C. Shi, "Graphic composite segmentation for PDF documents with complex layouts," Proc. SPIE 8658, Document Recognition and Retrieval XX, 86580E, February 2013.

[15] K. L. Chung, Y. H. Huang et. al. "Efficient sampling strategy and refinement strategy for randomized circle detection," Pattern Recognition, vol. 45, pp. 252-263, 2012

[16] K. Li, X. Lu, H. Ling, L. Liu, T. Feng, and Z. Tang, "Detection of overlapped quadrangles in plane geometry figures," in 12th International Conference on Document Analysis and Recognition., 2013.

[17] T. Feng, X. Lu, L. Liu, K. Li, and Z. Tang, "Structure analysis for plane geometry figures," in Document Recognition and Retrieval XXI(DRR 2014), Feb. 2014, in press.

[18] J. Song, X. Lu, H. Ling, X. Wang, Z. Tang , "Envelope extraction for composite shapes for shape retrieval," in 21st International Conference on Pattern Recognition.,2012.

[19] R. Sternberg, Cognitive Psychology, 3rd Edition, Thomson Wadsworth, 2003.

[20] K. Jarvelin, and J. Kekalainen, "IR evaluation methods for retrieving highly relevant documents," in Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 41-48, 2000.