QGAR – Querying Graphics through Analysis and Recognition

1 Team composition

Permanent researchers Benoît Naegel, MCF UHP until 12/ 2010; Philippe Dosch, MCF Nancy 2; Bart Lamirov, MCF INPL; Salvatore-Antoine Tabbone, PR Nancy 2; Laurent Wendling MCF until 06/2009.

Associate members Karl Tombre, DR INRIA, professor at INPL/Ecoles des Mines on secondment to INRIA, currently director of INRIA Nancy-Grand Est

Visiting scientist Makoto Hasegawa associate professor at Kinki University (Hiroshima, Japan), 2 months in 2010 and 7 months in 2011; Elisa Barney, associate professor at Boise State University (USA), 2 months in 2009; Djemel Ziou, professor at Université de Sherbrooke (Canada), 4 months (2007-09).

Postdocs, engineers Hervé Locteau, post-doc, funded by ScanPlan project (2009/11), UN2; Oriol Ramos-Terrades, ATER at UN2 (2007/08), associate professor UAB Barcelona since 2010; Vitor Vasconcelos, engineer under INRIA contract 2 years (2007/08).

Doctoral students Sabine Barrat, CIFRE Netlor Concept, defended in 2009, MCF Tours since 09/2010; Amani Boumaiza, Eureka Project 4462 ScanPlan, defense planned in 11/2012; Thanh Ha Do, MOET Vietnam and Eureka Project ScanPlan, defense planned in 2013; Mehdi Felhi, CIFRE OCE, defense planned in 2013; Salim Jouili, INRIA on ANR Navidomass project, defended in 2011, RD Architect Huawei Technologies (Belgium) since 05/2011; Santosh K.C, CORDI INRIA, defense planned in 2011; Oanh Nguyen, MOET Vietnam and Région-INRIA, defended in 2009, Associate professor at IPH Hanoi since 09/2010; Thai Hoang Van, BDI PED CNRS contract, defense planned in 12/2011; Jean-Pierre Salmon, European project FRESH, defended in 12/2008, Post-doc MICA Vietnam two years (2008-10)

2 Research topics

Keywords Document analysis, document indexing, pattern recognition, shape recognition, symbol spotting, performance evaluation

Research area and main goals The Qgar project-team works on the conversion of weakly structured information-an image of a paper document, or a PDF file, for example-into "enriched" information, structured in such a way that it can be directly handled within information systems. Our research belongs to the document analysis field, and more precisely to the graphics recognition community. We study the use of graphics recognition methods to index and organize weakly structured graphical information, contained in graphics-rich documents, such as technical documentation. In this context, we explore the capacity of pattern recognition methods to compute useful features for indexing and information retrieval.

3 Research activities

For many years, the main contributions of our team were in the area of algorithms and methods for image analysis and segmentation, with a specific focus on images of graphics-rich documents. In the last years, while keeping a regular activity in this domain, we have moved our main effort towards pattern recognition methods, especially for symbol recognition and spotting.
3.1 Document segmentation

In order to analyze a document several preprocessing steps are mandatory. The *binarisation* is usually the first step in most document image analysis systems. Binarisation methods have some well-known drawbacks such as for connections between distinct foreground objects or removal of small objects. Moreover, they are usually parameters dependent. An another step concerns the *Vectorisation*.

Existing methods generally suffer from two major drawbacks, over-segmentation and poor geometric precision, especially at the junctions between vectors.

*Text/graphic* separation is a hot problem in document image analysis. Extracting text is an important task since text has semantic meaning which could be obtained by a character recognition system. Extraction of text components is a challenging problem mainly because text components can vary in font styles and sizes and there may exist touchings either among text components or text and graphics components.

**Main results**  1) We proposed an original binarisation method based on connected operators which show good performance in various contexts and which is parameter free [NW10].

2) We proposed a set of robust vectorisation algorithms to detect extrema of curvature points and circular arcs [SW08b, JT08a, LG10] minimizing the previous drawbacks (see above).

3) Recently we investigate into sparse representation method. We proposed a robust method based on the MCA (Morphological Component Analysis) to separate the text from the graphic even if characters touch graphics [HT10b, HT10c].

**Self-assessment**  We made a lot of effort to propose efficient methods (binarisation and vectorisation) to segment documents. Even if these approaches are promising we believe that more and more applications of sparse representation will appear in the near future because they are more robust to segment complex documents (noisy and cluttered). Actually, in the field of document analysis and recognition, few sparse representation have been applied to digit recognition and text/segmentation from complex background. The results we obtain for the text/graphic separation recently [HT10b, HT10c] have high recall rate of text components, overcomes partially the problem of text/graphics touching, and outperforms the previous benchmark. We think that these approaches will be useful to segment complex documents in order to achieve symbols spotting.

3.2 Symbol recognition and spotting

Symbol recognition is the localization and identification of symbols in documents [NT10, TL07], to get natural features usable in indexing and retrieval applications. Symbol recognition still remains an open question when dealing with complex symbols having large variations or when their number is large and when symbols are embedded into the document. Our attention is focused on the weaknesses of the existing recognition methods, which make them difficult to be used.

**Main results**  1) *Statistical descriptors*: The Radon transform is a conversion of geometric transformations applied on a shape (for example a symbol) image into transformations in the radial and angular coordinates of the Radon image. Radon-based invariant shape descriptors [TRTB08] are different from the others in the sense that Radon transform is used as an intermediate representation upon which invariant features are extracted from for the purpose of indexing/matching. New directions for the utilization of the Radon transform for invariant shape representation have been explored and showed the efficiency and robustness of this transform to describe shapes combining the Radon, 1D Fourier-Mellin, and Fourier
transforms sequentially [HT10a] or generalizing the $\mathcal{R}$-signature [NTZH10c, NTZH10b].

2) **Structural descriptors**: These approaches deal, especially, with graph-based representations. Nevertheless, dealing with graphs suffers, on the one hand from the high complexity of the graph matching problem [LBHA+10] which is a problem of computing distances between graphs, and on the other hand from the robustness to structural noise which is a problem related to the capability to cope with structural variations and differences in the size of the graph. Firstly our attention was focused on the comparison of different graph similarity measures in the context of document retrieval [JTV10]. Secondly, we proposed a node signatures extraction combined with an optimal assignment method for approximating graph edit distance [JT09c]. We demonstrate that the graph edit distance formula can be written with an optimum solution of the assignment problem and that the proposed algorithm is robust for any type of graphs (labelled or not with symbolic or numeric labels).

Additionally, we have been focusing on extending work on spatial relation models [KWL10a] and have used it as a basis for complex symbol description, recognition and retrieval [KWL10b]. Our spatial relations handle fuzzy relations that convey a degree of truth rather than using standard all-or-nothing relations.

3) **Scaling of symbol recognition methods**: When handling a large dataset, a system needs an efficient index mechanism to retrieve data by their contents. In this perspective, we are interested in methods of information retrieval in masses of documents using structural representation [JT08b]. We propose a hypergraph structure [JT10b] where one graph can be assigned to more than one cluster. The structure enables to travel the data set and is efficient to cluster and retrieve graphs. The first results seem promising since we have efficiently indexed a database of 35000 molecules. In the context of graph clustering we propose an original algorithm based on an adaptation of the mean-shift [JTL10] and on the graph embedding method [JT10a]. This research took place in the scientific environment of the project Navidomass (Navigation In Document Masses, ANR MDCA) related to indexing large databases of ancient cultural heritage documents [COS+09, JCTO10].

4) **Symbol spotting**: As for the works above mentioned, symbol spotting concerns the retrieval of objects similar to a query within a database of images. We report in [NT10] a survey on symbol spotting and in the same vein we propose a method [NTB10, NTB09] based on visual vocabulary that is an original adaptation of information retrieval techniques. The local information is extracted in the surrounding of each point of interest. A document is "textualized" with a fuzzy matching technique to associate a point of interest with several visual words. Regions of interest in documents are identified by using local matching between the query and the documents. We use the vector model to compute the similarity between the regions of interest and the query symbol. Regions selected with high degree of similarity are considered as occurrences of the query symbol.

5) **Combination of classifiers and shape descriptors**: Combining outputs of classifiers, descriptors, or selecting features [CVCT09a, CRTT+08] is one of the strategies used to improve classification rates in common classification. We tackled the problem of combining classifiers within a non-Bayesian framework, considering both two-class and multi-class classifiers [RTVT09]. We are also interested in combining image descriptors and text modes using a probabilistic graphical models, especially Bayesian networks, to take into account different types of information and to manage missing data [BT10a, BT09a] or using Inductive Logic Programming to automatically learn non-trivial representations of the symbols [KLR09a, LR09].

**Self-assessment** The main efforts of the team have been focused on symbols recognition (and more generally on pattern recognition) and spotting. We are pioneer in the use of the Radon transform for shape recognition. In this direction, the R-transform, which gives rise
to the R-signature, is one of the most popular due to its simplicity. It has been successfully applied by others to several applications in the literature. Now we have reached robust results for isolated symbol recognition, especially with noisy background, and we need to extend and improve our methods to the recognition of symbols embedded into real documents (for instance symbol spotting). This topic is still challenging because of the appeal for preprocessing steps that introduce noise, heuristics to face computational complexity issue and we are faced to the paradox that to recognize symbols we need to segment the document and reciprocally. Even through the proposed approach on symbol spotting is only tested on synthetic documents, it gives interesting results and provides a possibility to apply on real documents.

There is a need in the community for scaling indexing methods. Our contribution on graph indexing is promising and in future works we want to improve this approach to deal with larger database (around several millions of graphs). In the same vein we also want to come back to the combination of classifiers or shape descriptors. When dealing with a large number of symbols, both signatures and structural recognition methods may not be powerful enough to discriminate. Combining outputs of classifiers or descriptors is one of the strategies used to improve recognition rates.

3.3 Performance evaluation and benchmarking

It is extremely important for the Document Image Analysis and Recognition community to be able to cross check and reproduce results described in published papers in the field. In order to achieve this, any datasets used as the basis for publications should be publicly available, as it is the norm in many other disciplines. Comparative studies should be also available so that a user facing a practical pattern recognition problem can get help in choosing the most appropriate family of descriptors. Since the end of 2004, our project-team is leader of the Epeires project affiliated to the Techno-Vision campaign. The objective is the construction of a complete environment for performance evaluation of symbol recognition and localization methods. This topic has gained increasing interest in the last years, as demonstrated by the creation of three international contests on symbol recognition methods [DVFE08].

**Main results** We lead the Epeires project\(^1\) on performance evaluation of symbol spotting and recognition (2005-2007) and, through this project we create testing data which have been used during evaluation campaigns, like the three editions of the International Contest on Symbol Recognition [VDW+07], organized during the last GREC workshops.

The web site is also the location where all resources related to the project are freely available to the scientific community and users are able to generate their specific testing evaluations.

We proposed several performance evaluation protocol [VRTT11, JT11, JTV10, JTV09, VTRTP07, ZDT07] for both pixel-based descriptors, computed on all pixels of the shape or on a subset of these pixels (contours, or regions, for example), and structural descriptors, computed from the components of the shape and from the relationships between them.

**Self-assessment** The Epeires web site is fully operational, and we continue to improve it, by adding new functionalities and by refining the existing content. We plan to use it to organize other incoming evaluation campaigns. Furthermore, we want to keep our fruitful collaboration with the CVC Barcelona on performance evaluation protocol. The performance evaluation we published are interesting and necessary for the community since few works exist in the literature in this direction. We can notice that the community uses the resources made available and that many papers still refer to it, as evidenced by the numerous references to reports on these evaluation campaigns.

\(^1\)http://www.epeires.org/
4 Scientific production

Number of publications

<table>
<thead>
<tr>
<th>Publication Type</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhD Thesis</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>International journal</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>International conference proceedings</td>
<td>10</td>
<td>6</td>
<td>15</td>
<td>12</td>
<td>2</td>
<td>45</td>
</tr>
<tr>
<td>National journal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>National conference proceedings</td>
<td>2</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Book or special issue (edited)</td>
<td>8</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td>13</td>
</tr>
</tbody>
</table>

Main publications  


Software, valorisation and technology transfert  

Since several years, the QGAR project-team has devoted much effort to the construction of a software environment, to be able to reuse whole or part of software implemented during previous work, as well as collected experience. The Qgar system is registered with the French agency for software protection (APP) and may be freely downloaded from its web site (http://www.qgar.org). The whole system is written in C++ and includes about 170,000 lines of code, including unit test procedures. A particular attention has been paid to the support of "standard" formats (PBM+, DXF, SVG), high-quality documentation, configuration facilities (using CMake), and support of Unix/Linux and Windows operating systems. The environment has been downloaded about 2.500 times since its availability, by academic users as well as industrial ones.

5 External supports and fundings

5.1 International cooperation

1) PAI Picasso with UAB Barcelona (2005-2007); 2) PAI Procore with city University Hong Kong (2005-2007); 3) PICS CNRS SEPIA with Mica Lab Hanoi (2007-2009); 4) STIC ASIE IDEA (2008-2010) with IFI Hanoi, IIT Hanoi, UKL Malaysia, L3I La Rochelle; 5) INRIA Euromed 3+3 AIDA (2009-2011) with the universities of: Annaba (Algeria), Rouen (France), Agadir (Morocco), Barcelona (UAB Barcelona, Spain) and Sousse (Tunisia).

5.2 European projects

1) Eureka SCANPLAN project (2009-2011) with UAB Barcelona, Anuman Interactive France, Icar Vision Systems (Spain); 2) Fresh project (2005-2007) with Algo ’tech Informatique (France), Estia (France), Euro Inter (France), Eads Sogerma Drawing (France), Ceit (Spain), Rector (Poland), Tekever (Portugal), and Zenon (Greece).

5.3 National projects

5.4 Industrial contracts


6 Collaborations

We have a long-lasting cooperation with the DAG (Document Analysis Group) from Universitat Autonoma Barcelona including joint PhD supervisions, publications, post-doc exchanges, joint organization of international symbol recognition contests. We were associated team (the topic was on performance evaluation) supported by INRIA from 2005 to 2008 and Université Nancy 2 in 2009. Through the ANR NAVIDOMASS we had a joint PhD (H. Chouaib visiting QGAR six months between 2007 and 2009) supervision with Université Paris 5 Descartes and several joint publications with Université de la Rochelle. Several collaborations with visiting scientists have given rise to several joint publications (some are submitted in 2011). Several local collaborations on vectorisation methods with the Adagio team and teams from CRAN (C. Join and V. Bombardier) lead to joint publications and on molecules graph indexing with ORPAILLEUR team (B. Maigret) from LORIA and with the Ecole Nationale Polytechnique d’Alger on image segmentation within a joint PhD, Nafaa Nacereddine, visiting QGAR from 10/2008 to 3/2010.

7 Teaching

Philippe Dosch is the director of studies for the bachelor degree ”Administration of open source systems, networks and applications”. Bart Lamiray headed the Department of Computer Science, and was the technical coordinator of the Ipiso specialized degree until 2009. Salvatore Tabbone heads one of the computer science masters (M2 Miage-ACSI) of Université de Nancy 2.

8 Visibility

Salvatore Tabbone is president of the GRCE (Groupe de Recherche en Communication Ecrite) since December 2010 and member of the editorial board of Journal of Universal Computer Science (JUCS). Salvatore Tabbone was the general chair of CIFED’08 and general chair of the steering committee of CIFED’10. Karl Tombre is editor in chief of the International Journal on Document Analysis and Recognition (IJDAR), member of the advisory board of Electronic Letters on Computer Vision and Image Analysis (ELCVIA), and member of the editorial board of Machine Graphics Vision and of Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées (ARIMA). Karl Tombre was past-president of the International Association for Pattern Recognition (IAPR) until august 2010. Salvatore Tabbone and Karl Tombre are members of the administrative council of AFRIF (French Association for Pattern Recognition and Interpretation). The member of the team were member of numerous committee programm of national and international conferences especially in their domain area: ICDAR’08-10, GBR’11, GREC’07-11, ICPR’08-10, DAS’08-10, ACM-SAC’07-11, CAIP’09-11, CIARP’08-10, RFIA’08-10, CIFED’08-12, CID’08-11, ORASIS’07-11, CART’08-10, TAIMA’07-11. The member of the team participated as reviewer to 11 committee PhD thesis, 3 to abroad and 1 HDR and as examinator for 11 PhD thesis and 5 HDR;
References


Thai V. Hoang, Salvatore Tabbone, and Ngoc-Yen Pham. Recognition-based segmentation of Nom characters from body text regions of stele images using area Voronoi diagram. In Xiaoyi Jiang and Nikolai Petkov, editors, International Conference on Computer Analysis of Images and Patterns - CAIP’2009, volume


