IAPR TC11 (Reading Systems)
Activity Report 2008-2010

Prof. Daniel Lopresti (Lehigh University, USA), TC11 Chair

Prof. Koichi Kise (Osaka Prefecture University, Japan), TC11 Vice Chair

http://www.iapr-tc11.org/

1. TC Background Information

1.1 Listing of TC Leadership Team

Toward our primary goal of supporting the community and its research activities, we have newly formed a TC-11 leadership team by adding three specialists, webmaster, dataset curator and newsletter editor, to a standard team consisting of the chair and the vice chair:

Role	Name	Affiliation	Email
Chair	Daniel Lopresti	Lehigh University, USA	lopresti@cse.lehigh.edu
Vice Chair	Koichi Kise	Osaka Prefecture University, Japan	kise@cs.osakafu-u.ac.jp
Webmaster	Masakazu Iwamura	Osaka Prefecture University, Japan	masa@cs.osakafu-u.ac.jp
Dataset Curator	Dimos Karatzas	Universitat Autónoma de Barcelona, Spain	dimos@cvc.uab.es
Newsletter Editor	Gernot Fink	TU Dortmund University, Germany	Gernot.Fink@tu-dortmund.de

The webmaster is responsible for the content of the TC-11 website, including helping to investigate new functionality to support the research community.

The dataset curator is responsible for managing the datasets on the website, including tracking down new datasets, investigating distribution issues, and providing additional annotation as needed.

The newsletter editor is responsible for working with the TC Chair to produce the monthly TC-11 newsletter, including soliciting information of interest to the community.

This is an experimental structure we have first introduced into TC11 and found that it works extremely well, since the burden is distributed to the members and their activities as specialists are far beyond the level that only the chair and the vice chair could achieve.

1.2 TC website URL

http://www.iapr-tc11.org/

1.3 Number of members (people on mailing list)

As of June 18th, 2010, there are exactly 1,538 subscribers to the TC-11 mailing list.

1.4 Communication types used (e.g. newsletters) and frequency

We mainly used two types of communication: web and newsletters. The frequency of the newsletter is monthly -- this has been dependable, regular like clock-work by the specialist, our newsletter editor.

1.5 Listing of key event(s) usually organised by the TC

TC11 has the following three key events held biannually:

	Event	Frequency	Topic area
ICDAR	International Conference on Document Analysis and Recognition	biannual, odd years	all fields on document analysis and recognition
ICFHR	International Conference on Frontiers in Handwriting Recognition	biannual, even years	handwriting recognition and its related areas
DAS	International Workshop on Document Analysis Systems	biannual, even years	document analysis methods and systems

In addition, TC11 is related to the following workshops which are held in conjunction with major events such as ICDAR:

	Event	Frequency	Topic area
AND	International Workshop on Analytics for Noisy Unstructured Text Data	annual	issues related to noisy text data by processing signals for human use
CBDAR	International Workshop on Camera-Based Document Analysis and Recognition	biannual, odd years	camera-based analysis of documents

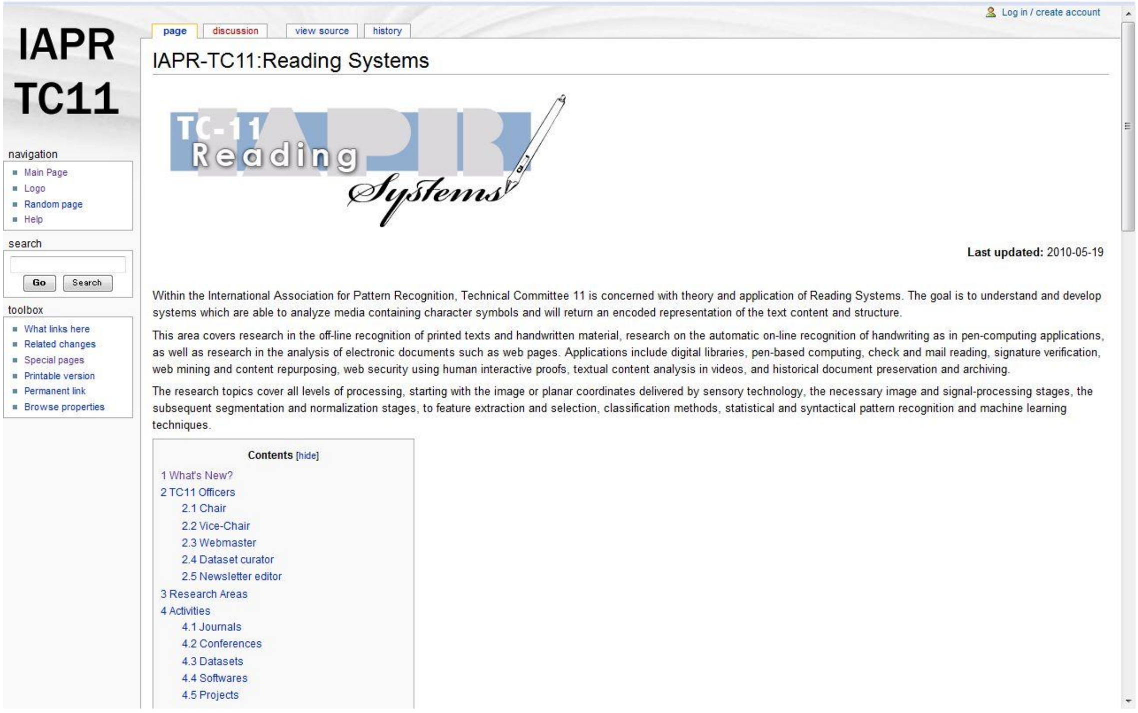
MOCR	International Workshop on Multilingual OCR	biannual, odd years	methodologies for multilingual document analysis systems with particular focus on OCR

2. Activities in the last two years (since ICPR2008)

We took over the duties of TC11 from Jianying Hu and Apostolos Antonacopoulos on February 15, 2009 and started our activities.

2.1 Website updates

The TC11 website was renovated in April 2010. The main purpose of the renovation was to make the website a web portal of our community. For the sake of that, we introduced a wiki system MediaWiki, developed for the Wikipedia, to keep the information up-to-date with an easy interface and obtain better look and feel as shown in the following figure.



IAPR-TC11:Reading Systems

Last updated: 2010-05-19

Within the International Association for Pattern Recognition, Technical Committee 11 is concerned with theory and application of Reading Systems. The goal is to understand and develop systems which are able to analyze media containing character symbols and will return an encoded representation of the text content and structure.

This area covers research in the off-line recognition of printed texts and handwritten material, research on the automatic on-line recognition of handwriting as in pen-computing applications, as well as research in the analysis of electronic documents such as web pages. Applications include digital libraries, pen-based computing, check and mail reading, signature verification, web mining and content repurposing, web security using human interactive proofs, textual content analysis in videos, and historical document preservation and archiving.

The research topics cover all levels of processing, starting with the image or planar coordinates delivered by sensory technology, the necessary image and signal-processing stages, the subsequent segmentation and normalization stages, to feature extraction and selection, classification methods, statistical and syntactical pattern recognition and machine learning techniques.

Contents [hide]

1 What's New?
2 TC11 Officers
2.1 Chair
2.2 Vice-Chair
2.3 Webmaster
2.4 Dataset curator
2.5 Newsletter editor
3 Research Areas
4 Activities
4.1 Journals
4.2 Conferences
4.3 Datasets
4.4 Softwares
4.5 Projects

2.1.1 Educational information

The primary responsibility of the TC-11 leadership team is supporting the international research community. In this sense, nearly everything we do has an educational component to it. By this we mean that students benefit as much as any other member of the community. In terms of activities that are specifically designed to encourage students and contribute to their training, the following is the list of items we have started to discuss (some of these are already in place, while others are new ideas which we may decide to implement):

(1) Tutorials

We have set up the link for ICDAR2009 tutorials video from the proceedings section:

http://www.iapr-tc11.org/mediawiki/index.php/Conferences#International_Conference_on_Document_Analysis_and_Recognition_.28ICDAR.29

(2) Summer school

(3) Best student paper awards at our conferences

We have already given the best student paper award and honorable mention at DAS2010. See http://www.cubs.buffalo.edu/DAS2010/ which is linked from the web of TC11.

(4) Support for lower student registration fees at our conferences

(5) Doctoral consortium

(see, e.g., http://iswc2010.semanticweb.org/node/18)

(6) Recorded lectures and/or interviews with founders of the field offering their advice to students:

It would be great if we could get each of senior leaders to record a message for future students with their best advice for success. This could be short interview which appears in each issue of the TC-11 newsletter.

(7) A contest or prize for getting students to sign up for the TC-11 mailing list

2.1.2 Tutorials

See "(1) Tutorials" of Sect. 2.1.1.

2.1.3 Description of application areas

- document processing: image-to-text
- check reading
- postal automation: envelope reading
- forms reading and parsing
- graphics recognition and beautification (Also see: TC-10)
- musical score reading
- mathematical equation reading
- signature verification
- pen computing

2.1.4 Examples of successful projects

We are thinking to include pointers and descriptions on projects and demos.

2.1.5 Demos

See the above Sect. 2.1.4.

2.1.6 Reference resources (datasets, evaluation tools)

One of the long-standing aims of the TC11 is to act as a point of reference for current datasets and practices in the community and track progress in Document Image Analysis research. In June 2009, the post of the "dataset curator" was created to address the above goal, and the following objectives were defined.

At the beginning of the term, the following objectives were defined:

O1. To organise the collection of datasets and related material.

O2. To enhance the user experience.

O3. To help making the TC11 Web site the portal site for document analysis.

O4. To ensure the durability and easy maintenance of the dataset collections in the future.

O5. To establish a link with TC5 and TC10 and avoid duplication of effort.

Over the past year, we have marked significant progress in each of the above targets, explained in detail in Appendix. Our activities have been focused on three axes. First, a discussion was initiated in the TC11, then extended to include our TC10 counterpart and finally was opened to the community during the 9th IAPR Workshop on Document Analysis Systems (DAS 2010). The aim was to get a better understanding of the needs and desires of the community. Second, we updated the existing presentation of datasets on the Web site and put in place a framework for receiving and hosting on the TC11 servers new material submitted. Third, with the help of the organisers of DAS 2010, we opened a call for new material and promoted intensively the issue of TC11 dataset services.

In 2010, we received expressions of interest to submit 7 new datasets to the TC11. Two of them are already online, three more are already in an advanced stage and are expected to be placed online within June – July 2010, while two are still in the expression of interest stage. In addition, we have completed substantial preparation work and defined a roadmap for the near future that we are confident that will result to the offering of a comprehensive service to the community. Appendix gives more details on this.

2.2 Research Initiatives

2.2.1 Events organised

The organized events are listed below. All events are fairly active as shown in the statistics of the number of presented papers as well as participants.

Event	Dates	Venue	Stats & URL
DAS2008	September 16-19, 2008	Nara, Japan	# papers: 80 # participants: 119 http://www.u-pat.org/das08/
ICDAR2009	July 26-29, 2009	Barcelona, Spain	# papers:277 # participants: 378 http://www.icdar2009.org/
DAS2010	June 9-11, 2010	Boston, USA	# papers: 65 # participants: 99 http://www.cubs.buffalo.edu/DAS2010/
ICFHR2010*	November 16-18, 2010	Kolkata, India	http://www.isical.ac.in/~icfhr2010/

*) upcoming conference

In addition to the above key events, we also have several important events some of which are endorsed by IAPR.

Event	Dates	Venue	URL, etc.
AND2008 endorsed by IAPR	July 24, 2008	Singapore	http://sites.google.com/site/and2008workshop/ in conjunction with ACM SIGIR 2008

AND2009 endorsed by IAPR	July 23-24, 2009	Barcelona, Spain	http://sites.google.com/site/and2009workshop/ in conjunction with ICDAR2009
CBDAR2009	July 25, 2009	Barcelona, Spain	https://sites.google.com/a/iupr.com/cbdar-2009/ in conjunction with ICDAR2009
MOCR2009	July 25, 2009	Barcelona, Spain	http://www.cubs.buffalo.edu/MOCR/

2.2.2 Publicity / dissemination activities

The activities for publicity and dissemination consists of the following 4 items:

(1) Newsletter

The TC11 newsletters consist of the following information:

- regular overview over most important conference-related dates (especially paper submission deadlines)
- on-demand-contents:
  - ➢ new/updated calls-for-papers for TC-11 related conferences and workshops
  - ➢ journal special issues/books
  - ➢ job opportunities
  - ➢ event reports (conferences, workshops)

In addition to the above regular contents, we provided the following new information:

- special content within 2009/2010
  - ➢ announcement of TC-11 logo contest and new logo/contest winner
- new regular content:
  - ➢ IJDAR Contents Telgram (introduced 4/2010 in cooperation with IJDAR Editor-in-Chief S. Marinai)

(2) Web

The TC11 web site provide visitors an up-to-date calendar of upcoming and past events of TC11 interests.

(3) Logo

We have defined a new logo of TC11 aiming at broader and easy recognition of our activities. We first called for a new logo to the TC11 members through the mailing list and received a submission by Eloisa Alquati. We have decided to accept her submission. The following is our new logo of TC11:

TC-11
IAPR
Reading
Systems

(4) Domain name

We have renewed the TC11 domain name (IAPR-TC11.ORG) for another 5 years (2009-2013).

2.2.3 Other

The TC-11 leadership played a role in determining the winners of the IAPR/ICDAR 2009 Awards. We also formed the ICDAR Advisory Board to help manage the bids and determine the location for ICDAR 2013, and determined that ICDAR2013 will be held in Washington DC, USA.

3. Plans (timeline until ICPR2012 and beyond)

We have the following plan until ICPR2012. The current important activities on the web page, newsletters and datasets are continued in the next two years. In addition, we start two new activities, software and summer school, for attracting new researchers and inviting them to the field of TC11.

- Web page: We update the webpage of TC11 frequently to attract researchers as a reliable source of information on reading systems. Expecially enriching the web contents by collecting educational materials and tutorials, finding successful projects and materials of nice demos are our primal target for the Web page.
- Newsletter: We continue to issue the newsletter monthly for anchoring the subscribers as well as obtaining new subscribers.
- Datasets: We consider that providing various datasets is one of the most important activities of TC11 in order to let new researchers come to the field

and do their research. Over the next period we plan to pilot a new infrastructure for submitting and using datasets and related material online, while we will keep an active role in archiving datasets used by our community.

- Software: Another important point for attracting new researchers to the field is to provide various freely available software. In addition to the datasets, we start collecting them and make them available at the TC11 web.
- Summer school: The TC11 team has just started exploring the idea of having a summer school just before or after a major conference/workshop such as DAS2012 for attracting more researchers. If there's enough interest, someone in the community will take charge of it. Our plan of the first step toward the summer school is to ask members of TC11 via the newsletter whether we should have the school, and if so what subjects they are interested in, etc.
- IAPR/ICDAR advisory board: The TC11 chair is a member of IAPR/ICDAR advisory board which will help ICDAR2011 organizers. It is of great importance since ICDAR2011 is the major event for the community after ICFHR2010. In addition to the activity as a member of the board, the chair of TC11 will act as one of the program chairs of ICDAR2011. The vice chair of TC11 will work as the workshops chair in ICDAR2011.

4 Recommendations to ExCo for TC leadership team for 2010-2012 term

We would like to propose to keep the same team for 2010-2012 term for continuing and extending the current activities of TC11.

Appendix

One of the long-standing aims of the TC11 is to act as a point of reference for current datasets and practices in the community and track progress in Document Image Analysis research. In June 2009, the post of the “dataset curator” was created to address the above goal. The dataset curator is responsible for collecting and managing useful resources for the community such as datasets, associated ground truth information and software. His responsibilities include tracking down new datasets, providing support and feedback to authors, appending small amounts of additional annotation as needed to make a collection useful to the community.

At the beginning of the term of the current team of TC11 Officers, in July 2009, we defined the following objectives in respect to the dataset management activities.

O1. To organise the collection of datasets and related material
  - To clean up the existing repository of datasets, remove obsolete ones and complete information where missing
  - To enrich the list of datasets by integrating other existing datasets, especially the ones related to competitions organised in the field over the past years
  - To establish a clear framework for contributing new datasets to TC11

O2. To enhance the user experience
  - To introduce a consistent way to present datasets on the Web site
  - To allow for easy searching and browsing through the collection
  - To implement ways to rank / rate material and solicit feedback from the users
  - To facilitate information exchange between users of the datasets

O3. To help making the TC11 Web site the portal site for document analysis
  - To promote the TC11 listed datasets as the basis for comparison between researchers
  - To encourage the inclusion of Ground Truth specification / information and Performance evaluation protocols along with submitted datasets

O4. To ensure the durability and easy maintenance of the dataset collection in the future
  - To implement a yearly “check and update” process

O5. To establish a link with TC5 and TC10 and avoid duplication of effort.

The dataset management related activities over the past year have been focused on three axes. First, a long discussion was initiated in the TC11, then extended to include our TC10 counterparts and finally was opened to the community during 9th IAPR Workshop on Document Analysis Systems (DAS 2010). The aim was to get a better understanding of the needs and desires of the community so that we can better address the above objectives. Second, with great help from the TC11 Webmaster, we updated the existing presentation of datasets and prepared the ground for receiving and hosting on the TC11 servers new material submitted. Third, with the help of the organisers of DAS 2010, we opened a call for new material and promoted intensively the issue of TC11 dataset services. As a result, the issue of datasets received a lot of attention during the workshop, with two special sessions organised on datasets, a keynote speech by Prof G. Nagy on the related topic of “Document Systems Analysis: Testing, Testing, Testing”, and a related working group. As a result of the promotion during DAS, we also received many new datasets that are currently in different stages of publishing online.

The progress registered under each of the objectives set in the beginning of our term is briefly summarised below.

O1. To organise the collection of datasets and related material. At the beginning of the term of the current team of TC11 Officers, the Web site of the TC11 listed links to 18 datasets under the thematic groups of "Machine Printed OCR" and "Handwriting Recognition". Only 3 datasets could be considered to be complete and up to date, while for the majority of the datasets the Web site listed non-working links and incomplete information. The list of material has since been updated and old, unsupported datasets have been removed. In June 2010, the TC11 lists 11 datasets.

We have been active in soliciting new datasets, and created a framework to deal with new submissions. During the past year this was achieved through the initiative organised around DAS 2010, were authors were encouraged to submit data referred to in their papers and special sessions on datasets were organised. In the context of the DAS2010 initiative, we received expressions of interest to submit 7 new datasets. Two of them are already online, and contain associated ground truth information and visualisation software as well. Three more are already in an advanced stage and are expected to be placed online within June – July 2010, while two more are still in the expression of interest stage.

An important aspect that we have looked into extensively is the right framework for soliciting submissions. One of the important issues identified early on was ensuring that the authors have copyright ownership over the data submitted. Ways to safeguard TC11 from any potential copyright infringement issues have been discussed in length. During the DAS 2010 workgroup discussion we identified the Digital Millennium Copyright Act (DMCA, USA) and the Electronic Commerce Directive (ECD, EU) as good platforms to base the TC's activities (limiting the liability of providers of on-line services for copyright infringement by their users). Uploading datasets and related material will be managed within the context of the above frameworks, while the use of a creative commons license to the content will be actively encouraged.

O2. To enhance the user experience. In the short term, and with the help of the TC11 Webmaster, we have updated the look and feel of the datasets representation on the Web site, and have defined a minimum set of information that we request from authors with every new submission, so that future categorisation and searching is easier (an example of the new template can be seen at the new submissions received during DAS 2010).

In the medium to longer term, we would like to look into a number of ways to enhance the user experience. This entails further changes to the TC11 Web site, but also changes in the infrastructure. We have already defined and explained a number of such improvements, including ranking and rating of submitted material, facilities to leave feedback and initiate discussions on specific datasets etc. These are explained in more detail in intermediate TC11 reports and working documents during the past year (available through the TC11 dataset curator).

In terms of infrastructure, we have identified a platform that would allow us extensive functionality in storing and retrieving data ("Data Analysis and Exploitation" project: http://dae.cse.lehigh.edu/DAE/). During the following term it is planned to test the suitability of this platform in collaboration with TC10.

O3. To help making the TC11 Web site the portal site for document analysis. Over the past year we have actively promoted the TC11 Web site as the prime location to publish new datasets. During DAS 2010 and with the help of the organisers, we actively encouraged authors to publish the data they used for testing their algorithms. The initiative was welcomed by the community, which demonstrates that there is willingness to make data available if the right infrastructure is in place. This has worked well for the TC11, as it served as a reminder that the TC11 provides such a service. It is our opinion that by tapping on the built momentum it is possible to largely achieve this objective during the next term.

O4. To ensure the durability and easy maintenance of the dataset collection in the future. Previous experience has shown that datasets quickly get out of date (which is probably a good thing, as we would avoid over-training on the same data). One of the big challenges is to deal with out-of-date information, and dependencies on data that are not available anymore. Towards this we have worked in two directions. First, in terms of hosting material, we had a long debate regarding whether data should be hosted at the TC11 site, or should be allowed to be hosted externally. While we still support both ways we encourage authors to host their data with the TC11 (largely solving the problem of dead-links and versioning issues). Second, in organisations that rely on volunteer work like the IAPR and the TC11 in particular, it is important to reduce the maintenance overhead for managing submitted material. We expect that the new platform we are planning to introduce in the near future will simplify maintenance tasks, and allow the dataset curator to assume a more appropriate coordinating rather than technical role.

O5. To establish a link with TC5 and TC10 and avoid duplication of effort. Over the past year we have established a link and are working closely with the TC10, and the TC10 dataset curator Dr Bart Lamiroy. The two committees have a common understanding on the community needs, and we are planning to create a common Web portal for the two TCs.

During the past year, the issue of collecting useful material (datasets, ground-truth information, software etc) and making it available to the community has received a lot of attention within TC11. The progress made is important, but even more so is the fact that we now have a clearer idea of what is needed by the community, and what can be achieved within the context and resources of TC11. We are confident that the roadmap we have produced will result to the offering of a comprehensive service to the community, and we are looking forward to the future of this initiative.