IAPR TC11 (Reading Systems)
Activity Report 2010-2011

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 goal of supporting the community and its research activities, we continue with the same successful TC-11 leadership team we established two years ago:

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

Beyond the standard roles of Chair and Vice Chair, 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 Chair to produce the monthly TC-11 newsletter, including soliciting information of interest to the community.

This structure, which we introduced into TC-11 when we assumed leadership in 2009, has proven to be extremely effective and allows us to accomplish goals beyond what the chair and the vice chair could achieve on their own.

1.2 TC website URL

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

1.3 Number of members (people on mailing list)

As of June 20th, 2011, there are 1,537 subscribers on the TC-11 mailing list.

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

We mainly used two types of communication: the TC-11 website and a monthly newsletter. The frequency of the newsletter has been extremely dependable over the past two years due to the diligent oversight of our Newsletter Editor. In addition to the monthly mass emailing, past newsletters are archived on the TC-11 website.

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

TC-11 has oversight responsibility for the following three major events:

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, TC-11 members are involved in organizing the following workshops, most of which are typically held in conjunction with ICDAR (* = endorsed by IAPR):

Event		Frequency	Topic area
AFHA*	International Workshop on Automated Forensic Handwriting Analysis	TBD	Handwriting analysis, signature verification, and handwriting forensics
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	biannual, odd years	camera-based analysis of documents

	Analysis and Recognition		
HIP	International Workshop on Historical Document Imaging and Processing	TBD	digital imaging, collection and processing of historical documents
MOCR	International Workshop on Multilingual OCR	biannual, odd years	methodologies for multilingual document analysis systems with particular focus on OCR

2. Activities over the past year

We assumed leadership of TC-11 in February 2009. This report will largely focus on activities that have taken place over the past year (i.e., after our last Activity Report), with additional details as needed to provide proper background. The most significant items of note include updates to the TC-11 website, continued growth in our data collection activities, and the creation of a first-time ICDAR Doctoral Consortium.

2.1 Website updates

The TC11 website was significantly revised and updated in April 2010. The main purpose for this change was to make the website a flexible, powerful portal for our community. We introduced a wiki system, MediaWiki, developed for Wikipedia, to facilitate keeping information up-to-date with an easy interface and provide a better look-an- feel as shown below. The Webmaster ensures that the website is updated frequently to keep the content fresh and relevant.

Major sections of the website include:

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

o 4.5 Projects

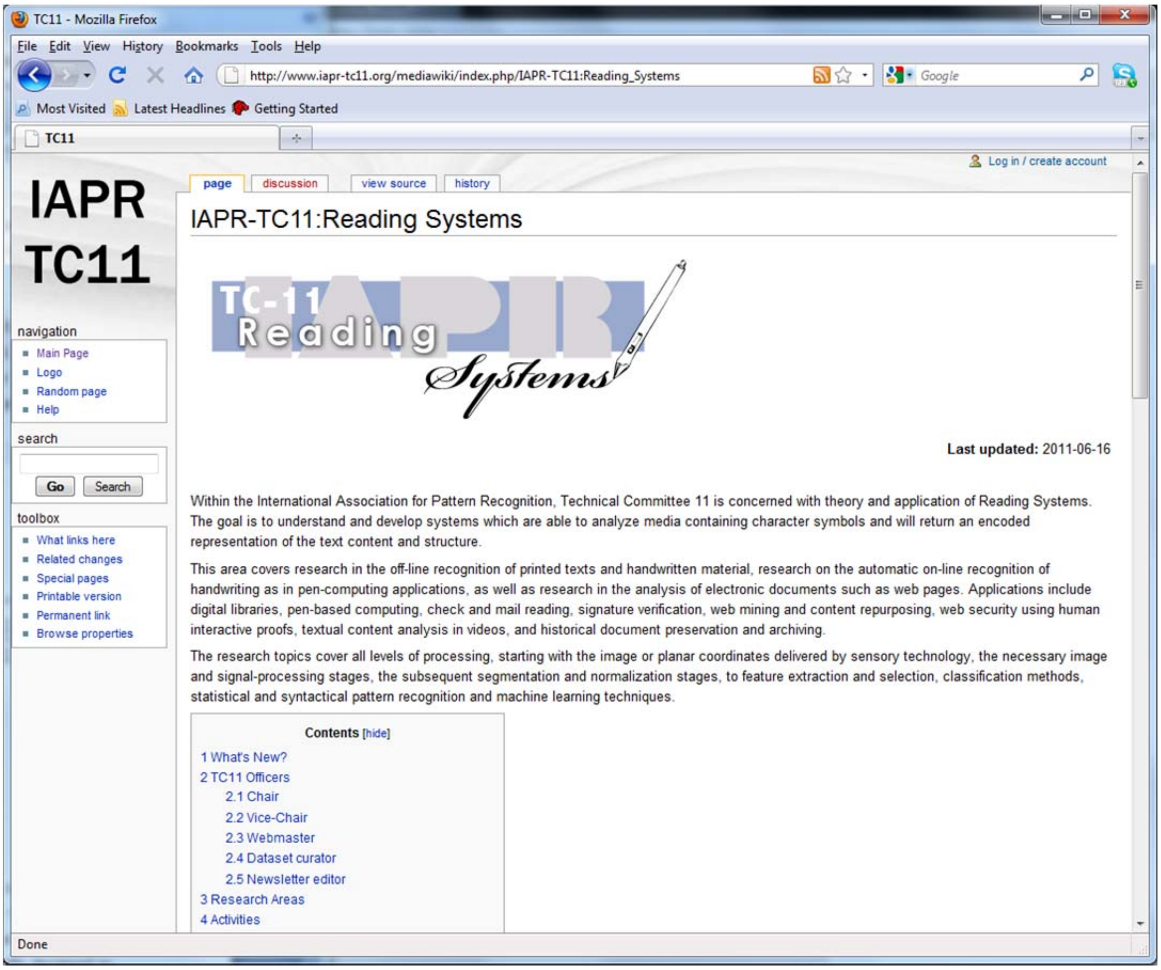
o 4.6 Forums

* 5 Newsletters

o 5.1 Newsletter Archive

o 5.2 How to receive the IAPR TC-11 newletters?

* 6 Logo



IAPR-TC11:Reading Systems

Last updated: 2011-06-16

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]

A notable recent addition to the TC-11 website is an archive of past conference websites: http://www.iapr-tc11.org/archive/. We consider this to be an extremely valuable service to the community. Frequently websites for events become unavailable (go "dead") after the conference is over and are lost to today's researchers. We have recently started offering to archive conference websites on the TC-11 web site. The material we archive also includes publications when these appear on the main conference web site (i.e., when an agreement with the publisher is in place, as, for example, in the case of ICDAR 2009).

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 most significant new development we have to report this year is the establishment of the first ICDAR Doctoral Consortium, to be held in conjunction with the ICDAR conference in Beijing in September 2011.

The concept of a Doctoral Consortium is popular in other research communities as a way of supporting young researchers and strengthening the field, and through the ICDAR 2011 event, we will explore the benefits for the international document analysis community. The goal of the Doctoral Consortium is to create an opportunity for Ph.D. students to test their research ideas, present their current progress and future plans, and receive constructive criticism and insights related to their future work and career perspectives. A mentor (a senior researcher who is active in the field) will be assigned to each student to provide individual feedback. In addition, students will have the opportunity to present an overview of their research plan during a special poster session.

The leaders of TC-11 conceived the idea of a Doctoral Consortium and sought feedback through consultation with the ICDAR Advisory Board and the ICDAR 2011 organizers, and by conducting a survey of the community via the TC-11 mailing list. After determining the feasibility of the idea, we invited the leaders of TC-10 to join us in co-organizing the event (the ICDAR conference is co-sponsored by TC-11 and TC-10). Further details concerning the Doctoral Consortium are available on the ICDAR 2011 website: http://www.icdar2011.org/EN/column/column39.shtml.

In addition to the significant addition of the ICDAR Doctoral Consortium, the following activities are either currently supported or under consideration for the future:

(1) Best student paper awards at our conferences. It is typically the case that TC-11 conferences offer an IAPR Best Student Paper Award.

(2) Support for lower student conference registration fees and affordable hotel options.

(3) Online resources on the TC-11 website, including pointers to past conferences and proceedings. For example, we have set up a link for ICDAR 2009 tutorials videos.

Discussion is ongoing in two areas where we may attempt experiments later this year:

(4) A summer school for document analysis.

(5) Recorded lectures and/or interviews with founders of the field offering their advice to beginning researchers. A number of leaders in pattern recognition and document analysis have retired recently or are on the verge of doing so, and we plan to explore providing video messages from them on the TC-11 website.

2.1.2 Tutorials

As noted above, we provide links to TC-11 conferences, some of which have made their tutorials available online (e.g., ICDAR 2009).

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

Postal address reading and check processing are two application areas now in commercial use. Optical character recognition and limited forms of higher level document analysis are available in off-the-shelf software products.

2.1.5 Demos

See the above Sect. 2.1.4.

2.1.6 Reference resources (datasets, evaluation tools)

An important service offered by TC-11 is the possibility for authors to archive datasets along with ground truth data and software on the Web site of the TC. The availability of good datasets in the public domain is recognized as an important factor for boosting innovation and tracking progress in a research field. The document analysis field has traditionally lacked behind in the provision of freely available datasets and associated evaluation tools. This is for good reasons, including the fact that in most of the cases large collections of documents include private information, which makes its distribution impossible. TC-11 recognizes the need to inflict a change of mentality, and is actively working in promoting the dissemination of such information by inviting submissions, offering advice and providing archiving space through its web site.

The objectives set out for the role of the dataset curator (see previous report for more details) are the following:

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 TC-5 and TC-10 and avoid duplication of effort.

Over the past year we have marked progress in each of the above targets. The main activities are summarized below.

New Datasets

We have archived or linked to numerous new datasets and software tools through the website of TC-11, and disseminated information of new datasets through the TC-11 newsletter.

Resource Name	Status
Machine Print Document Images	
Table Ground Truth for the UW3 and UNLV datasets	Archived
The DocLab Dataset for Evaluating Table Interpretation Methods	Archived
PRImA Layout Analysis Dataset NEW	Linked
DFKI Dewarping Contest Dataset (CBDAR 2007) NEW	Linked
APTI: Arabic Printed Text Image Database NEW	Linked
Graphical Document Images	
Chem-Infty Dataset: A ground-truthed dataset of Chemical Structure Images	Archived
Scene Text Images	
KAIST Scene Text Database NEW	Archived
Mixed-Content Document Images	
Tobacco800 Document Image Database NEW	Linked
On-line and Off-line Handwriting	

CASIA Online and Offline Chinese Handwriting Databases NEW	Linked
On-line Handwriting	
Devanagari Character Dataset	Archived
Harbin Institute of Technology Opening Recognition Corpus for Chinese Characters (HIT-OR3C)	Archived
IAM Online Document Database (IAMonDo-database)	Archived
IAM On-Line Handwriting Database	Linked
UNIPEN database	Linked
Nakagawa Lab Online Handwriting Database NEW	Linked
The Informal Competition of Recognizing On-line Words (ICROW)	Linked
Off-line Handwriting	
IBN SINA: A database for research on processing and understanding of Arabic manuscripts images	Archived
CEDAR Off-line Handwriting CDROM1	Linked
IAM Database	Linked
The GERMANA Dataset NEW	Linked
The RODRIGO Dataset NEW	Linked
MARG- Medical Article Records Groundtruth NEW	Linked
Hindi font samples	Linked
Software and Tools	
GEDI: Groundtruthing Environment for Document Images NEW	Linked
Submission in Progress	
ICDAR 2003/2005 Robust Reading Competitions	Transferring
CD Covers Dataset	Transferring

Dataset Sourcing Activities

We have been active in soliciting new datasets. In that respect we have updated the submission process, given comments we received and experience we accumulated over the first year of our term. The new submission form is simpler and more intuitive to fill-in and to understand.

We have taken an active role in inviting submissions. In particular, we contacted all the organizers of past competitions in the field (usually in the context of the ICDAR conference series), and requested that they archive their datasets with TC-11. We had limited success with this initiative for numerous reasons. The main issues reported to us are listed below, and provide the guidelines for future actions by TC-11.

A few authors were not willing to archive their datasets to TC-11, as that way they would lose control of who is accessing them. In some cases this has to do with an unwillingness to offer datasets for free to commercial entities, and the desire by authors to restrict access to academics for research purposes only. This is difficult to manage (copyright notices do not seem to persuade authors, see below). Especially as far as monetary compensation for non-academic use is concerned, the TC is neither willing (as a matter of principle) nor able to offer such infrastructure. Nevertheless, it is a possibility to add support for monitoring download activity. Actually, this is linked to one of our initiatives: ranking of the datasets online. This is a technical issue that is related to the web site infrastructure, as it would require users to create accounts and log in. We are looking into this.

The second most important aspect is copyright issues. Understandably this is a grey area for most academics. As part of the submission process we request the authors to verify that they own the copyright to the resources they submit for online archiving. If not able to do so, we ask them to explain us who owns the copyright and why they think they are able to place the resource in the public domain (e.g., the material is already publicly available under a creative-commons license). For the case of competitions we did not expect this to be any problem, given that the datasets had already been made available in the past for the very purposes of hosting the competition. Nevertheless, there were authors that were taken aback by this copyright message. More often than not this reflected a lack of knowledge about copyright situations, but in any case, it acted as a deterrent for authors.

Our standard practice as far as copyright is concerned is to advise authors to use some of the creative commons copyright statements when submitting owned work. Nevertheless, we are not able, nor are we qualified, to offer any further legal advice on the subject. Especially in the case of document datasets, this has proven to be a continual problem.

One way forward that we are currently considering is to reformulate the submission process so that the authors can directly upload data on our site, instead of submitting it to the dataset curator first. Moving to this submission model would place us under the Digital Millenium Copyright Act (USA) and the EU Copyright Directive, passing the responsibility of uploading materials to the users instead of the TC-11 (see the operation framework of YouTube, Facebook, Twitter etc). This is both a technical issue (re-purposing the web site) and a logistic one (currently the hosting for TC-11 datasets is done in Japan which does not fall under either of the above directives).

User Interface Aspects

Changing the user interface and providing new functionalities is a key objective, although technically demanding to advance on voluntary time. Apart from the aspects discussed above, which are still in the planning, we have introduced changes in the web site that we hope would make it easier to use.

Apart from the categorisation of resources according to research topic, we have introduced a sorting based on the journal / conference where they were first announced. We consider it important for the authors to perceive added value from archiving resources with TC-11, and one way to achieve this is from receiving citations to related papers. Organising the datasets according to the journal/conference where they first appeared could promote citations. In addition, we believe that explicitly showing the link between the dataset and the publication will help in connecting the act of publishing with making the dataset available, hence facilitate the shift of mentality.

Collaborations

As set out in the original objectives, we are working closely with TC-10 in order to avoid duplicated efforts and streamline the management of the dataset submission process. We have now agreed on a common submission process (through a single form), and have coordinated the invitations for new datasets. We are considering initiating an online forum on the datasets topic.

2.2 Research Initiatives

2.2.1 Events organised

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

Event	Dates	Venue	Stats & URL
DAS 2008	September 16-19, 2008	Nara, Japan	# papers: 80 # participants: 119 http://www.u-pat.org/das08/
ICDAR 2009	July 26-29, 2009	Barcelona, Spain	# papers: 277 # participants: 378 http://www.icdar2009.org/
DAS 2010	June 9-11, 2010	Boston, USA	# papers: 65 # participants: 99 http://www.cubs.buffalo.edu/DAS2010/

ICFHR 2010	November 16-18, 2010	Kolkata, India	# papers: 117 # participants: 129 http://www.isical.ac.in/~icfhr2010/
ICDAR 2011	September 18-21 2011	Beijing, China	# papers: 278 (as of June 20, 2011) # participants: ? http://www.icdar2011.org/
DAS 2012	March 27-29	Gold Coast, Australia	http://www.ict.griffith.edu.au/das2012/
ICFHR 2012	September 18-20	Bari, Italy	http://www.icfhr2012.uniba.it/
ICDAR 2013	TBD	Washington, DC	

In addition to the above major events, TC-11 is also directly or indirectly responsible for a number of workshops through its active members (* = IAPR endorsed):

Event	Dates	Venue	URL, etc.
AND 2008*	July 24, 2008	Singapore	http://sites.google.com/site/and2008workshop/ in conjunction with ACM SIGIR 2008
AND 2009*	July 23-24, 2009	Barcelona, Spain	http://sites.google.com/site/and2009workshop/ in conjunction with ICDAR 2009
CBDAR 2009	July 25, 2009	Barcelona, Spain	https://sites.google.com/a/iupr.com/cbdar-2009/ in conjunction with ICDAR 2009
MOCR 2009	July 25, 2009	Barcelona, Spain	http://www.cubs.buffalo.edu/MOCR/ in conjunction with ICDAR 2009
AND 2010*	October 26, 2010	Toronto, Canada	http://sites.google.com/site/and2010workshop/ in conjunction with CIKM 2010
AFHA 2011*	September 17-18, 2011	Beijing, China	http://forensic.to/webhome/afha/ in conjunction with ICDAR 2011
AND 2011*	September 17, 2011	Beijing, China	http://and2011.cse.lehigh.edu/ in conjunction with ICDAR 2011
CBDAR 2011	September	Beijing,	http://imlab.jp/cbdar2011/

	22, 2011	China	in conjunction with ICDAR 2011
HIP 2011	September 16-17, 2011	Beijing, China	http://www.comp.nus.edu.sg/~hdocp/ in conjunction with ICDAR 2011
MOCR 2011	September 17, 2011	Beijing, China	http://www.cubs.buffalo.edu/MOCR/ in conjunction with ICDAR 2011
ICDAR Doctoral Consortium	September 18, 2011	Beijing, China	http://www.icdar2011.org/EN/column/column39.shtml

2.2.2 Publicity / dissemination activities

The activities for publicity and dissemination consist of the following:

(1) TC-11 monthly newsletter

The TC-11 newsletters consist of a regular overview of most important conference-related dates (especially paper submission deadlines), new/updated calls-for-papers for TC-11 related conferences and workshops, journal special issues/books, Table of Contents for the latest IJDAR issue, job opportunities, and event reports (conferences, workshops). The newsletter is sent to over 1,500 researchers who are members of our mailing list.

(2) TC-11 web site

The TC-11 web site provides visitors an up-to-date calendar of upcoming and past events of TC-11 interest. In addition, the website is also the portal by which users can access datasets we have collected and our archive of past conference websites. The TC-11 domain name (IAPR-TC11.ORG) is renewed until 2013.

2.2.3 Other

The TC-11 leadership plays a role in the IAPR/ICDAR Awards process. The chair of TC-11 also forms and serves on the ICDAR Advisory Board to help manage the bids and determine the location for future ICDAR conferences. The location for ICDAR 2015 will be decided by the community at a joint TC-11 / TC-10 meeting at the ICDAR 2011 conference in Beijing. In addition, we also are involved in managing the bidding process for future ICFHR conferences as well. (Bidding to host the DAS workshop series is handled through a different process.)

3. Plans (timeline until ICPR 2012 and beyond)

We believe we have made excellent progress over the past year and a half. We will continue with the activities we have put into place, namely our ongoing efforts on dataset collection and improvement of the TC-11 website, and organizing the first-ever ICDAR Doctoral Consortium. In addition, we plan to explore the idea of a document analysis summer school and recording videos of leading researchers in the field for dissemination on the TC-11 website. As always, we will remain open and on the lookout for new opportunities to serve the TC-11 community as well as to collaborate with colleagues form other IAPR technical committees.