HIP 2015 Program

8:30 Welcoming by Bertrand Coüasnon, Chair HIP'15

Session 1: Text Transcription (Chair: Liangrui Peng)

8:45	Bastien Moysset, et. al.	Space Displacement Localization Neural Networks to locate origin points of handwritten text lines in historical documents
9:05	Konstantinos Zagoris, et. al.	A Framework for Efficient Transcription of Historical Documents Using Keyword Spotting
9:25	Alejandro Toselli, et. al.	Handwritten Text Recognition Results on the Bentham Collection with Improved Classical N-Gram-HMM methods
9:45	Thomas Packer, et. al.	Cost-Effective Information Extraction from Lists in OCRed Historical Documents
10:05	George Retsinas, et. al.	Historical Typewritten Document Recognition Using Minimal User Interaction
10:25	Stefan Pletschacher, et. al.	Europeana Newspapers OCR Workflow Evaluation

10:45 Break

Session 2: Segmentation and Layout Analysis (Chair: Apostolos Antonacopoulos)

11:00	Maroua Mehri, et. al.	Learning Texture features for Enhancement and Segmentation of Historical Document Images
11:20	Hao Wei, et. al.	Selecting Autoencoder Features for Layout Analysis of Historical Documents
11:40	Boqiang Fan, et. al.	Layout Analysis Algorithm Based on Probabilistic Graphical Model for Dunhuang Historical Documents
12:00	Joan Pastor-Pellicer, et. al.	Combining Learned Script Points and Combinatorial Optimization for Text Line Extraction
12:20	Muhammad Zeshan Afzal, et. al.	Document Image Binarization using LSTM: A Sequence Learning Approach
12:40	Mathias Seuret, et. al.	Clustering Historical Documents Based on the Reconstruction Error of Autoencoders

13:00 Lunch

Session 3: Templates, Date Estimation, and Script Specific Approaches (Chair: Volker Märgner)

14:15	Elisa H. Barney Smith, et. al.	Template generation from postmarks using cascaded unsupervised learning
14:35	Yuanpeng Li, et. al.	Publication Date Estimation for Printed Historical Documents using Convolutional Neural Networks
14:55	Fredrik Wahlberg, et. al.	Large scale style based dating of medieval manuscripts
15:15	Bartosz Bogacz, et. al.	Homogenization of 2D & 3D Document Formats for Cuneiform Script Analysis
15:35	Nicholas Howe, et. al.	A Character Style Library for Syriac Manuscripts
15:55	Leonard Rothacker, et. al.	Retrieving Cuneiform Structures in a Segmentation-free Word Spotting Framework

16:15 Award of the IAPR Best Paper

16:20 Closing by Bill Barrett, HIP General Chair

16:30 Break

16:50 Walk to the Library of Nancy

Abstracts

Session 1: Text Transcription

8:45 - 9:05
Space Displacement Localization Neural Networks to locate origin points of handwritten text lines in historical documents
Bastien Moysset, Pierre Adam, Christian Wolf and Jérôme Louradour

We describe a new method for detecting and localizing multiple objects in an image using context aware deep neural networks. Common architectures either proceed locally per pixel-wise sliding-windows, or globally by predicting object localizations for a full image. We improve on this by training a semi-local model to detect and localize objects inside a large image region, which covers an object or a part of it. Context knowledge is integrated, combining multiple predictions for different regions through a spatial context layer modeled as an LSTM network.

The proposed method is applied to a complex problem in historical document image analysis, where we show that is capable of robustly detecting text lines in the images from the ANDAR-TL competition. Experiments indicate that the model can cope with difficult situations and reach the state of the art in Vision such as other deep models.

9:05 - 9:25
A Framework for Efficient Transcription of Historical Documents Using Keyword Spotting
Konstantinos Zagoris, Ioannis Pratikakis and Basilis Gatos

Keyword spotting (KWS) has drawn the attention of the research community as the alternative means to solve hard cases of handwriting text recognition. In this paper, a framework is proposed that employs KWS to enhance the efficiency in the manual transcription process, thus, reducing drastically the cost of training data creation. The core principle relies upon the ability of robust document-specific descriptors to produce meaningful similarities between a chosen word image for transcription and the corresponding word images in the full dataset under consideration. In the proposed framework, KWS is coupled with a relevance feedback mechanism, which further enhances retrieval performance while being independent to the chosen KWS algorithm. The efficiency of the proposed pipeline is showcased via a user-friendly web-based prototype.

9:25 - 9:45
Handwritten Text Recognition Results on the Bentham Collection with Improved Classical N-Gram-HMM methods
Alejandro Héctor Toselli and Enrique Vidal

Handwritten Text Recognition experiments and results are presented on the historical Bentham text image dataset used in the ICFHR-2014 HTRtS competition. The successful segmentation-free holistic framework is adopted, using traditional modelling approaches based on Hidden Markov (HMM) optical character models and an N -gram language model (LM). Departing from the very basic N -gram-HMM base-line system provided in HTRtS, several improvements are made in text image preprocessing, feature extraction and LM and HMM modeling, including more accurate HMM training by means of discriminative training. As a result, we achieve similar word recognition accuracy as some of the best performing (single, uncombined) systems based on (recurrent) Neural Networks (NN), under the same training and testing conditions. Using the traditional N -gram/HMM framework has several advantages over modern approaches based on (hybrid, recurrent) NNs. Perhaps the most important are the dramatically faster training of HMMs and the high stability of HMM Baum-Welch training results with respect to model initialization. Thanks to this stability, HMM-based systems can be easily trained with just one arbitrary initialization, thereby making it generally unnecessary to try several (or many) random initializations or to resort to other bootstrapping and/or randomization strategies, as often needed for NN training. These advantages become crucial when dealing with many historical document collections, which are typically huge and entail very high degrees of variability, making it generally difficult to re-use models trained on previous collections.

9:45 - 10:05
Cost-Effective Information Extraction from Lists in OCRed Historical Documents
Thomas Packer and David W. Embley

To work well, machine-learning-based approaches to information extraction and ontology population often require a large number of manually selected and annotated examples. In this paper, we propose ListReader which provides a way to train the structure and parameters of a Hidden Markov Model (HMM) without requiring any labeled training data. This HMM is capable of recognizing lists of records in text documents and associating subsets of identical fields across related record templates and is particularly well-suited for information extraction from OCRed historical documents. The algorithmic training method we employ is based on a novel unsupervised active grammar-induction framework that, after producing an HMM wrapper, uses an efficient active sampling process to complete the mapping from the HMM wrapper to ontology by requesting annotations from a user for automatically-selected examples. We measure performance of the final HMM in terms of F-measure of extracted information and manual annotation cost and show that ListReader learns faster and better than a state-of-the-art baseline (CRF) and an alternate version of ListReader that induces a regular expression wrapper.

10:05 - 10:25
Historical Typewritten Document Recognition Using Minimal User Interaction
George Retsinas, Basilis Gatos, Apostolos Antonacopoulos, Georgios Louloudis and Nikolaos Stamatopoulos

Recognition of low-quality historical typewritten documents can still be considered as a challenging and difficult task due to several issues i.e. the existence of faint and degraded characters, stains, tears, punch holes etc. In this paper, we exploit the unique characteristics of historical typewritten documents in order to propose an efficient recognition methodology that requires minimum user interaction. It is based on a pre-processing stage in order to enhance the quality and extract connected components, on a semi-supervised clustering for detecting the most representative character samples and on a segmentation-free recognition stage based on a template matching and cross-correlation technique. Experimental results prove that even with minimum user interaction, the proposed method can lead to promising accuracy results.

10:25 - 10:45
Europeana Newspapers OCR Workflow Evaluation
Stefan Pletschacher, Christian Clausner and Apostolos Antonacopoulos

This paper summarises the final performance evaluation results of the OCR workflow which was employed for large-scale production in the Europeana Newspapers project. It gives a detailed overview of how the involved software performed on a representative dataset of historical newspaper pages (for which ground truth was created) with regard to general text accuracy as well as layout-related factors which have an impact on how the material can be used in specific use scenarios. Specific types of errors are examined and evaluated in order to identify possible improvements related to the employed document image analysis and recognition methods. Moreover, alternatives to the standard production workflow are assessed to determine future directions and give advice on best practice related to OCR projects.

Session 2: Segmentation and Layout Analysis

11:00 - 11:20
Learning Texture Features for Enhancement and Segmentation of Historical Document Images
Maroua Mehri, Nibal Nayef, Pierre Héroux, Petra Gomez-Krämer and Rémy Mullot

Many challenges and open issues related to the tremendous growth in digitizing collections of cultural heritage documents have been raised, such as information retrieval in digital libraries or analyzing page content of historical books. Recently, graphic/text segmentation in historical documents poses specific challenges due to many particularities of historical document images (e.g. noise and degradation, presence of handwriting, overlapping layouts, great variability of page layout). To cope with those challenges, a method based on learning texture features for historical document image enhancement and segmentation is proposed in this article. The proposed method is based on using the simple linear iterative clustering (SLIC) superpixels, Gabor descriptors and support vector machine (SVM) model. It has been evaluated on 100 document images which were selected for historical document layout analysis and historical book recognition competitions in the context of ICDAR conference and HIP workshop (2011 and 2013). To demonstrate the enhancement and segmentation quality, the evaluation is based on manually labeled ground truth and shows the effectiveness of the proposed method through qualitative and numerical experiments. The proposed method provides interesting results on historical document images having various page layouts and different typographical and graphical properties.

11:20 - 11:40
Selecting Autoencoder Features for Layout Analysis of Historical Documents
Hao Wei, Mathias Seuret, Kai Chen, Andreas Fischer, Marcus Liwicki, Rolf Ingold and Xiuqin Zhong

Automatic layout analysis of historical documents has to cope with a large number of different scripts, writing supports, and digitalization qualities. Under these conditions, the design of robust features for machine learning is a highly challenging task. We use convolutional autoencoders to learn features from the images. In order to increase the classification accuracy and to reduce the feature dimension, in this paper we propose a novel feature selection method. The method cascades adapted versions of two conventional methods. Compared to three conventional methods and our previous work, the proposed method achieves a higher classification accuracy in most cases, while maintaining low feature dimension. In addition, we find that a significant number of autoencoder features are redundant or irrelevant for the classification, and we give our explanations. To the best of our knowledge, this paper is one of the first investigations in the field of image processing on the detection of redundancy and irrelevance of autoencoder features using feature selection.

11:40 - 12:00
Layout Analysis Algorithm Based on Probabilistic Graphical Model for Dunhuang Historical Documents
Boqiang Fan, Liangrui Peng and Franck Lebourgeois

The Dunhuang historical documents are of great significance to the study of ancient Chinese Buddhist culture and other topics. It would greatly benefit the protection and the study of historical documents with full-text information generated by historical document recognition technology. However, many historical documents from Dunhuang are old and broken, and to make it worse, the style and layout of these documents are casual as well. Traditional layout analysis algorithm failed to pay much attention to these problems. In this paper, a new layout analysis algorithm based on Probabilistic Graphical Model is proposed, including both rough segmentation and fine segmentation. After the input historical document images are pre-processed by Gaussian smoothed filtering and binarization, the rough segmentation step uses projection information to get rough text-column regions. In the fine segmentation step, a connected component analysis algorithm based on Probabilistic Graphical Model is developed. The method models the extracted connected components based on Markov Random Field, and combines connected components to get output text columns. Experiments were conducted on some Dunhuang historical documents, and the proposed method could correctly segment text columns with a recall rate of 90.0% and an accuracy of 77.7%. The segmented text-column regions could cover 99.2% characters in historical document images. The result shows that the proposed layout analysis algorithm could be successfully applied to degraded historical document images.

12:00 - 12:20
Combining Learned Script Points and Combinatorial Optimization for Text Line Extraction
Joan Pastor-Pellicer, Garz Angelika, Rolf Ingold and Maria Jose Castro-Bleda

Complex layouts, curved text lines, heterogeneous background, noise, and clutter still render text line extraction in the context of historical documents a challenging task where traditional methods do not excel. We propose a novel text line extraction method with two contributions: first, text-specific interest points extracted by supervised machine learning; and second, reformulating the problem of bottom-up text line aggregation as noise-robust combinatorial optimization. In a final step, unsupervised clustering eliminates invalid text lines. Building the method on top of interest points and posing aggregation as global optimization problem, we can detect text lines with arbitrary orientation and curvature, and are robust to noise and clutter. Experimental evaluations on the IAM Saint Gall dataset show promising results.

12:20 - 12:40
Document Image Binarization using LSTM: A Sequence Learning Approach
Muhammad Zeshan Afzal, Joan Pastor-Pellicer, Faisal Shafait, Thomas Breuel, Andreas Dengel and Marcus Liwicki

We propose to address the problem of Document Image Binarization (DIB) using Long Short-Term Memory (LSTM) which is specialized in processing very long sequences. Thus, the image is considered as a 2D sequence of pixels and in accordance to this a 2D LSTM is employed for the classification of each pixel as text or background. The proposed approach processes the information using local context and then propagates the information globally in order to achieve better visual coherence. The method is robust against most of the document artifacts. We show that with a very simple network without any feature extraction and with limited amount of data the proposed approach works reasonably well for the DIBCO 2013 dataset. Furthermore a synthetic dataset is considered to measure the performance of the proposed approach with both binarization and OCR groundtruth. The proposed approach significantly outperforms standard binarization approaches both for F-Measure and OCR accuracy with the availability of enough training samples.

12:40 - 13:00
Clustering Historical Documents Based on the Reconstruction Error of Autoencoders
Mathias Seuret, Andreas Fischer, Angelika Garz, Marcus Liwicki and Rolf Ingold

The term "historical documents" encompasses an enormous variety of document types considering different scripts, languages, writing supports, and degradation degrees. For automatic processing with machine learning and pattern recognition methods, it would be ideal to share labeled learning samples and trained statistical models across similar documents, avoiding a retraining from scratch for every historical document anew. In this paper, we propose to cluster historical manuscripts based on the reconstruction error of autoencoders. A low reconstruction error suggests visual similarity between a new manuscript and a known manuscript, for which the autoencoder was trained in an unsupervised fashion. Preliminary experiments conducted on 10 different manuscripts written with ink on parchment demonstrate the ability of the reconstruction error to group similar writing styles. For discriminating between Carolingian and cursive script, in particular, near-perfect results are reported.

Session 3: Templates, Date Estimation, and Script Specific Approaches

14:15-14:35
Template Generation from Postmarks using Cascaded Unsupervised Learning
Elisa H. Barney Smith and Gernot Fink

A method for automatically extracting templates for each category of these postmark stamps is described. The problem is complicated by the high levels of degradation present in the cards. The ink is faded, and the paper is yellowed. A significant quantity of both occlusion and dropout exists in the same images. The postmark stamps overlap the other content on the card. The rubber stamp sometimes did not make full contact with the paper resulting in an incomplete image. Excesses of ink connect the text and line components. The approach uses a cascade of unsupervised learning steps separated with image cleaning. The templates once extracted can be used to group the postmarks, and will contribute information about the postmark content to better separate the postmark from the paper and other interfering marks to extract further information about the postmarks and postcards.

14:35-14:55
Publication Date Estimation for Printed Historical Documents using Convolutional Neural Networks
Yuanpeng Li, Dmitriy Genzel, Yasuhisa Fujii and Ashok Popat

This paper describes an approach to estimating the unknown publication date for printed historical documents from their scanned page images, using Convolutional Neural Networks. The method primarily harnesses visual features from small image patches. Optionally, we augment the feature set with textual OCR result features to improve accuracy, though at greater preprocessing cost. To be applied in various tasks, we develop both classification and regression models. As an example application, we show that Optical Character Recognition (OCR) can be improved if we use estimated publication date to select the appropriate OCR model. Moreover, the resulting improvement in OCR accuracy is close to what could be achieved knowing the true publication date. We are not aware of previous work in estimating publication dates for printed historical documents with visual features.

14:55-15:15
Large scale style based dating of medieval manuscripts
Fredrik Wahlberg, Lasse Mårtensson and Anders Brun

In this paper we propose a novel approach for manuscript dating based on shape statistics. Our goal was to develop a strategy well suited for a large scale dating effort where heterogeneous collections of thousands of manuscripts could be automatically processed. The proposed method takes the gray scale image as input. Uses the stroke width transform and a statistical model of the gradient image to find ink boundaries. Finally, a distribution over common shapes, quantified using shape context descriptors, is produced for each manuscript image. The proposed method is binarization free, rotational invariant and requires minimal segmentation. Also, we propose parameters estimation schemes, simplifying the deployment process.
We evaluate our work on the 10000+ manuscripts collection "Svenskt diplomatariums huvudkarotek", consisting of charters from the medieval period of todays Sweden. The images, originally intended for web viewing, were of low quality and had compression artifacts, but could still be dated with a median absolute error of < 19 years. Due to unsupervised feature learning and regression, we get these results using only 5% of the labels in the estimator training.

15:15-15:35
Homogenization of 2D & 3D Document Formats for Cuneiform Script Analysis
Bartosz Bogacz, Judith Massa and Hubert Mara

In the Digital Humanities, text sources can be digitized using various methods resulting in different data representations of related documents. This challenge is exacerbated in for clay tablets with cuneiform script, which is one of the oldest handwritten scripts used for more than three millennia. A cuneiform tablet acquired using a 3D-Scanner and a manually created line tracing are two completely different representations of the same type of text source. Additional a line tracing can be born-digital as vector graphic or it can be a raster image of a drawing with ink on paper. Each representation is typically processed with its own tool-set and therefore the textual analysis is limited to a certain type of digital representation. In this work we present a work-flow for unification of the three most common graphical representations of cuneiform tablets. The first approaches vectorizes the manually created retro-digitized tracings by skeletonization and applies pattern matching to extract the wedges, which are the radical elements of cuneiform script. Secondly, the born-digital drawings also require pattern matching as the curved lines are set differently by each draftsperson. Due to the density of wedges a subsequent conflict resolution is applied to both types of line drawings. As cuneiform script is a handwriting in 3D, we show the segmentation and extraction of wedges from high-resolution 3D-models. The result is one representation exported as Scalable Vector Graphic (SVG), which is used for character retrieval for a future Optical Character Recognition (OCR) as ultimate goal.

15:35-15:55
A Character Style Library for Syriac Manuscripts
Nicholas Howe, Alice Yang and Michael Penn

Paleographers study ancient and historical handwriting in order to learn more about documents of significant interest and about their creators. Computational tools and methods can aid this task in numerous ways, particularly for languages and scripts that are not widely known today. One project currently underway seeks to gather a collection of securely dated letter samples from Syriac documents dating between 500 and 1100 CE. The set comprises over 60,000 human-selected character samples. This paper gives details on the collection and describes the automatic techniques used to process the initial human input so as to produce high-quality segmented character samples ready for analysis.

15:55-16:15
Retrieving Cuneiform Structures in a Segmentation-free Word Spotting Framework
Leonard Rothacker, Denis Fisseler, Gerfrid G.W. Müller, Frank Weichert and Gernot A. Fink

Cuneiform tablets are an invaluable documentation of early human history. Efforts are being made in digitizing large tablet collections for preserving their content and making them available to a global research community. However, there are hardly any automated computer aided methods for supporting philologists in their analysis. In this paper we present an approach for automatically retrieving cuneiform wedge constellations from digitized cuneiform tablet collections. Encouraging results could be achieved in our qualitative and quantitative evaluation on a challenging benchmark consisting of 3D-scanned cuneiform tablets.