DAS 2016 Tutorial proposal:
Scene-Text Localization, Recognition, and Understanding

Presenters:

Albert Gordo
Xerox Research Center Europe
albert.gordo@xrce.xerox.com

Lluís Gómez i Bigordà
Computer Vision Center, Universitat Autonòma de Barcelona
lgomez@cvc.uab.es

[Invited talk, unconfirmed]
Max Jaderberg
Google DeepMind
max.jaderberg@gmail.com

Abstract:

During the last few years, the computer vision and document analysis communities have started giving attention to tasks related to text localization and recognition in natural images (also referred to as scene-text or text-in-the-wild), particularly after the seminal works of Wang et al. More recently, and steered by the current deep learning renaissance, architectures based on convolutional neural networks and recurrent neural networks have shown outstanding results on localization and recognition tasks, and have allowed researchers to approach more challenging problems such as text *understanding* in natural images.

This tutorial has three main objectives: first, **to familiarize the audience with the problem of localization, recognition, and understanding of text in natural images**, highlighting the similarities and differences between them and the same tasks performed on document images. Second, **to provide some details about the techniques that are showing the largest potential for current and future research in the topic**, and that could be easily transferred or adapted back to the document analysis domain. Third, **to present to the audience open-source libraries that implement some of the current state-of-the-art methods**.

Intended Audience:

The intended audience are researchers interested in tasks related to text localization, recognition, and understanding in images. These type of tasks are key to construct robust reading systems for text understanding in unconstrained scenarios.

The tutorial will focus on natural images, although many of the ideas presented here can be extrapolated and adapted to document images, particularly in cases involving complex documents such as mixed content documents, figures, charts, etc.

The material should be suitable for all types of researchers, from PhD students and post-docs to senior researchers. Some basic knowledge about common computer vision and

machine learning techniques (e.g. traditional local features for image representation such as SIFT or HOG, image encodings such as bag of words or Fisher vectors, supervised classification with support vector machines, energy minimization with conditional random fields, etc) is highly encouraged, but will not be necessary to get the main messages of the tutorial.

Similarly, basic knowledge about deep learning (particularly convolutional neural networks (CNNs) and recurrent neural networks such as long short-term memory networks (LSTMs) is encouraged but not necessary.

Expected Audience:

The topic is very interesting for document image analysis and the wider computer vision community. As an indication, the Robust Reading Competition (RRC) has more than 1000 registered users (covering all continents and practically all countries in Europe), many of which from the document analysis field. The RRC has received 200,000 page views and many thousands of downloads of the datasets which is an indication of the interest. In 2014 the 1st Int. W. on Robust Reading was organised as a satellite event of ACCV 2014. It was only of the better attended events with a top attendance of around 50 attendees. As another indication of interest within the community, IJDAR published an Special Issue in 2014 on Robust Reading.

Apart from these direct indications of relevance to the community, the topics of these workshop cover a lot of state-of-the-art methodologies that might be of interest to the DAS audience independently of the particular context of the tutorial.

These indications make us believe that most of the registered participants at DAS could have a potential interest to participate.

Motivation and learning outcomes:

Text localization, text recognition, and text understanding are core tasks that appear naturally in most document analysis problems and that have been explored for several decades. However, dealing with these tasks in natural images has only recently started to receive attention, mostly after the seminal works of Wang et al [Wang2010; Wang2011].

Despite the similarities and the potential transferability of techniques between the document and the natural image domains, a large amount of the literature related to text localization and recognition in the wild has been published outside of the more traditional channels of the document analysis community. As a result, it is easy to be unaware of the large advances that have been produced in the topic, as well as of the recent techniques that have shown outstanding results and that could be easily transferred or adapted back to the document domain.

For these reasons, we believe a tutorial on scene-text tasks can be of high interest to the document analysis community.

The goals of this tutorial are as follows:

- To introduce the attendees to the problems of text localization, text recognition, and text understanding in natural images, highlighting the similarities and differences between them and the same tasks performed on document images.
- To provide a brief historical overview of the main approaches used to address these problems.
- To discuss the current state-of-the-art techniques to address these problems, and how these techniques, which might be unknown to the attendees, can be leveraged in the document analysis domain.
- To provide the attendees with open-source tools, code, and models to easily replicate some of the state-of-the-art results and to adapt them to document analysis tasks.

Topic and description:

Scene text understanding consists in determining whether a given image contains textual information and if so, localizing it and recognizing its written content.
Typically, this is achieved through a two-stage pipeline: first, regions where text may appear are localized in the image. Then, the text inside those regions is recognized.

Text localization:

The large number of techniques proposed for text localization in natural scenes can be divided into patch-based, region-based, and hybrid approaches. Patch-based methods usually work by performing a sliding window search over the image and extracting certain features (e.g. HOG [Wang2010; Mishra2012], or convolutional features [Coates2011; Jaderberg2014a]) in order to classify each possible patch as text or non-text. On the other hand, region-based methods [Ephstein2010; Neumann2012; Gomez2013; Yin2014] are based on a typical bottom-up pipeline: first performing an image segmentation to generate a set of character candidates, filtering the resulting regions, and finally leading a grouping process where characters candidates are grouped together into words or text lines.

More recently, after the success of generic region proposals for object localization [vanDeSande2011; Zitnick2014], Jaderberg et al explored the possibility of using generic region proposals for the task of text localization [Jaderberg2015]. After filtering these proposals with a text/non text classifier and improving the regions using bounding box regression, this approach obtained excellent recall and precision results on standard datasets.

Text recognition:

These localization methods are able to provide regions that potentially contain text. However, in most cases, they do not directly provide the transcription of the text: **the recognition has to be performed as an independent step**.

Most current methods are based on three different strategies:

- **Explicitly localizing and recognizing the individual characters of the word**, similar in spirit to optical character recognition (OCR). The work of Bissacco et al [Bissacco2013] is a recent example.
- **Implicitly localizing and recognizing the individual characters jointly**. Early works such as [Mishra2012] are based on handcrafted features such as HOG and energy minimization using e.g. conditional random fields, while more recent approaches [Shi2015] can learn the features and the models in an end-to-end manner using convolutional neural networks jointly with recurrent neural networks such as LSTMs.
- **Explicitly constructing a global representation of the word image, and learning classifiers for the individual characters or the whole word using the global representation.** Similar to the previous case, early works are based on constructing global representations with handcrafted standard features such as SIFT aggregated with bags of words or Fisher vectors [Almazan2014; Rodriguez2015; Gordo2015a] and where the learning is done with “shallow” methods (support vector machines, metric learning). More recent works such as [Jaderberg2014b] learn both the image representation and the classifiers jointly using convolutional neural networks.

Currently, the last two strategies are the ones obtaining the best results in terms of recognition accuracy, mainly due to the use of convolutional neural networks. It is worth noting, however, that the third approach has the added advantage of producing a global signature of the word image, which allows one to perform other related tasks such as word image retrieval, indexing, clustering, etc.

Beyond text recognition:

Traditionally, works dealing with text recognition and retrieval have focused on the task of retrieving or recognizing *exactly* the same word, without any considerations of the semantics of the word. However, for practical purposes, one may be interested in having representations of word images that encode not the characters it contains, but its meaning. This would allow one to perform more challenging tasks leading to document or scene understanding.

To the best of our knowledge, this task has not been extensively addressed by either the document analysis or the computer vision communities until very recently (the only exception we are aware of is the work of Krishnan and Jawahar [Krishnan13]). We believe one of the main reasons for this is that, with the typical features used until very recently, these types of tasks seemed unattainable. However, given the recent advances in word image representation and recognition, this task seems now feasible.

The very recent LEWIS [Gordo2015b] addresses this exact problem by learning how to embed word images and semantic concepts in a joint space and where images that share the same meaning are encoded closely in space. We believe this type of semantic representations is necessary to pave the road towards full understanding of content containing textual information.

Program:

The tutorial will be split in three different sessions of approximately one hour each, covering each of the three previous topics: text localization, text recognition, and text understanding.

Each session will contain a theoretical part, where the most relevant works will be introduced, and a practical part, where we will show how to use open-source libraries and models to achieve the results discussed in the theoretical part.

Text Localization (50 minutes + 10 minutes for questions)

***Presented by Lluís* Gómez**

In this session we will cover a basic introduction to scene text localization techniques. Starting from a clear definition of the problem and its evaluation protocols, we will make a quick review of the evolution of the state-of-the-art highlighting the most relevant methods. The session will have an important practical component where we will present the OpenCV text module [OpenCVText], an open source project providing different algorithms for text detection and recognition in natural scene images. We will provide different demo programs written in Python with which the participants will be able to perform text localization on their own images.

Theoretical part (20 minutes approximately)

We plan to cover the following topics:

- Introduction to scene text localization. Definition of the problem and why we need it. How it differs from generic object detection. How has been typically approached.
- Evaluation protocols. Common issues with ground truth and detection granularities.
- State of the art review. By looking at the evolution of results in the ICDAR Robust Reading Competition dataset. We'll link to public implementations of methods when available and go into details for the most relevant ones.
- Open challenges. More unconstrained environments: Incidental text, multi-language, arbitrary orientations.

Practical part (30 minutes approximately)

For the practical part, we will first introduce the OpenCV text module. Then we will provide code samples to replicate some recent methods on text localization

- Demo: Detecting English horizontal text with Class Specific Extremal Regions [Neumann2012] and Exhaustive Search grouping [Neumann2011].
- Demo: Detecting language-independent arbitrary-oriented text with MSER [Matas2004] and similarity hierarchies [Gomez2014].
- Demo: A simple Text Proposals algorithm for word spotting in the wild [Gomez2015].

Text Recognition (50 minutes + 10 minutes for questions)

Presented by Max Jaderberg if available, otherwise by Albert Gordo

In this session we will cover the problem of text recognition: given a cropped image of a word, how can we recognize its contents? We will provide a quick review of the methods typically used for this task, which can be divided in three groups: i) methods that localize and classify the individual characters of the word image explicitly, ii) methods that implicitly localize the characters and jointly recognize the characters and the word, and iii) methods that represent the word image with global features and earn character / word classifiers. As previously discussed, learning an image representation offers other advantages, such as being able to perform word image matching.

Theoretical part (30 minutes approximately)

We plan to cover the following topics:

- Introduction to the problem
- Description of the three main families of approaches, with brief comments on the most relevant works of each family
- Description, in some more detail, of the recent deep learning methods that are currently achieving state-of-the-art results
- Highlight the connection with text recognition in document analysis and the transferability potential of the state-of-the-art techniques

Practical part (20 minutes approximately)

In the practical part, we will show how to use the pretrained models of Jaderberg et al. [Jaderberg2014b] to perform word image recognition from Python using the popular Caffe library. We will also use the pretrained models to extract word image features that can be used for other tasks such as word image retrieval.

If time allows it, we will also show how to train new models to other tasks more related to document analysis such as handwritten word recognition.

Text Understanding (40 minutes + 10 minutes for questions)

Presented by Albert Gordo

In this part of the tutorial we will first present some recent works that show the importance of semantics in text images (e.g. [Movshovits2015]). We then will show some recent works from the text analysis community showing how semantic representations can be learned for text [Mikolov2013]. Finally we will introduce LEWIS [Gordo2015b], an approach to embed word images and semantic concepts in a joint space that leverages the latest advances in word recognition, and where images that share the same meaning are encoded closely in space.

Theoretical part (25 minutes approximately)

We plan to cover the following topics:

- Importance of semantics in text and scene understanding
- Distributional and non-distributional semantics embeddings for for text words
- Semantic embeddings of word images: LEWIS

Practical part (15 minutes approximately)

In the practical part, we will show how to use the LEWIS pretrained models to extract semantic features from word images, and show how these features can be used to retrieve semantic concepts or other images that share the same semantics.

As in the previous case, these models rely on the Caffe library and can be used from Python.

Previous editions:

This is the first time this tutorial will be given.

Material:

We will provide the following material to the attendees:

- Course slides
- Code and pre-trained models to reproduce some of the experiments shown during the tutorial. The main code will be in Python, and rely on libraries written in C or C++

Some of the libraries used in the tutorial (e.g. OpenCV, Caffe) are not always easy to install. We will try to provide a virtual machine image with the needed libraries already installed.

Equipment

No extra equipment will be required.

Biographies

Albert Gordo *is a research scientist at the Computer Vision group at Xerox Research Center Europe (XRCE) in Grenoble, France. Before that, after finishing his PhD in the Document Analysis Group at the Computer Vision Center, Universitat Autònoma de Barcelona, he spent one year as a postdoctoral researcher at the LEAR group at INRIA Grenoble, France.*

One of his main research interests is text recognition in images, with several works published in top conferences and journals (CVPR, ICCV, PR, TPAMI, IJCV), and where one of these works won the Handwritten Keyword Spotting *competition at ICFHR 2014. Currently, he is particularly interested in tasks related to text understanding in images, i.e., going beyond simple text recognition.*

Lluís Gómez i Bigordà *received the B.Sc. and M.Sc. degrees in Computer Science from Universitat Oberta de Catalunya. He obtained M.Sc. degree in Computer Vision and Artificial Intelligence in 2010 from Universitat Autònoma de Barcelona, where he is currently a PhD Candidate and a research assistant in the Computer Vision Center within the Document Analysis and Pattern Recognition Group.*

His research interests are on computer vision and machine learning techniques applied to scene text understanding. As a member of the document analysis community he has had the chance to collaborate with a variety of research groups and venues. He has done a research stay at the Osaka Prefecture University in Japan, and has collaborated with other prominent research groups in the organization of the ICDAR Robust Reading

Competition in their 2013 and 2015 editions. In 2015 he served as a member of the Program Committee in the 6th International Workshop on Camera Based Document Analysis and Recognition (CBDAR). Lluís is an enthusiast of Free Open Source Software with more than 15 years of experience in software development. He has contributed to several Open Source projects, e.g. the Pure Data visual programming language and the OpenCV computer vision library among others. In 2013 and 2014 he was selected for participation in two consecutive editions of the Google Summer of Code program.

***Max Jaderberg** is a research scientist at Google DeepMind in machine learning and computer vision. Previously co-founded Vision Factory which was acquired by Google in 2014, and completed his PhD at the Visual Geometry Group, University of Oxford under the supervision of Prof. Andrew Zisserman and Prof. Andrea Vedaldi.*

During his PhD, Max Jaderberg explored the use of deep learning for tasks related to text spotting, creating systems that have very significantly advanced the state-of-the-art in the topic. His main interests are in artificial intelligence, deep learning, and perception.

Minimal bibliography:

[OpenCVText] http://docs.opencv.org/3.0-beta/modules/text/doc/text.html

[Almazan2014] J. Almazan, A. Gordo, A. Fornes, and E. Valveny, "Word spotting and recognition with embedded attributes". TPAMI, 2014.

[Bissacco2013] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven, "PhotoOCR: Reading Text in Uncontrolled Conditions". In ICCV, 2013.

[Coates2011] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. Wu, and A. Ng, "Text detection and character recognition in scene images with unsupervised feature learning". In ICDAR, 2011.

[Epshtein2010] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform". In CVPR 2010.

[Gomez2013] L. Gomez and D. Karatzas, "Multi-script text extraction from natural scenes" in Proc. ICDAR, 2013.

[Gomez2014] L. Gomez and D. Karatzas, "A fast hierarchical method for multi-script and arbitrary oriented scene text extraction". CoRR abs/1407.7504, 2014.

[Gomez2015] L. Gomez and D. Karatzas, "Object Proposals for text extraction in the wild", In ICDAR 2015.

[Gordo2015a] A. Gordo, "Supervised Mid-level features for Word Image Representation". In CVPR, 2015.

[Gordo2015b] A. Gordo, J. Almazan, N. Murray, and F. Perronnin, "LEWIS: Latent Embeddings for Word Images and their Semantics". In ICCV 2015.

[Jaderberg2014a] M. Jaderberg, A. Vedaldi, A. Zisserman, "Deep features for text spotting". In ECCV 2014.

[Jaderberg2014b] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman. "Synthetic data and artificial neural networks for natural scene text recognition". In NIPS DLW 2014.

[Jaderberg2015] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, "Reading Text in the Wild with Convolutional Neural Networks". IJCV 2015.

[Krishnan2013] P. Krishnan and C.V. Jawahar, "Bringing Semantics in Word Image Retrieval". In ICDAR 2013.

[Matas2004] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions". IVC 2004.

[Mikolov2014] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. "Distributed Representations of Words and Phrases and their Compositionality". In NIPS 2013.

[Mishra2012] A. Mishra, K. Alahari, and C. Jawahar, "Top-down and bottom-up cues for scene text recognition". In CVPR 2012.

[Movshovitz2015] Y. Movshovitz-Attias, Q. Yu, M. C. Stumpe, V. Shet, S. Arnoud, & L. Yatziv, "Ontological Supervision for Fine Grained Classification of Street View Storefronts". In CVPR 2015.

[Neumann2011] L. Neumann and J. Matas, "Text Localization in Real-world Images using Efficiently Pruned Exhaustive Search". In ICDAR 2011.

[Neumann2012] L. Neumann and J. Matas, "Real-Time Scene Text Localization and Recognition". In CVPR 2012.

[Rodriguez2015] J.A. Rodriguez, A. Gordo, and F. Perronnin, "Label Embedding: A Frugal Baseline for Text Recognition". IJCV 2015.

[Shi2015] B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-based Sequence Recognition and its Application to Scene Text Recognition". CoRR abs/1507.05717, 2015.

[vanDeSande2011] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. Smeulders, "Segmentation as selective search for object recognition". In ICCV 2011.

[Wang2010] K. Wang and S. Belongie, "Word spotting in the wild". In ECCV 2010.

[Wang2011] K. Wang, B. Babenko, and S. Belongie, "End to End Scene Text Recognition". In ICCV 2011.

[Yin2014] X.C. Yin, X. Yin, K. Huang, H.W. Hao, "Robust text detection in natural scene images". TPAMI 2014.

[Zitnick2014] C. L. Zitnick, and P. Dollár, "Edge boxes: Locating object proposals from edges". In ECCV 2014.