CLEF-IP Image Tasks Guidelines

May 30, 2011

1 Introduction

The CLEF–IP track was launched in 2009 to investigate IR techniques for patent retrieval and was part of the CLEF 2009 evaluation campaign. In 2011, the track continues as a benchmarking activity of the CLEF 2011 conference.

There are five tasks in the 2011 track:

- Prior Art Candidate Search: Find patent documents that are likely to constitute prior art to a given patent application.
- Classification: Classify a given patent document according to the IPC system, up to the subclass level.
- New: Refined Classification: Classify a given patent document up to the group/subgroup level, when the subclass is given.
- New: Image-based Document Retrieval: Find patent documents relevant to a given patent document containing images.
- New: Image-based Classification: Categorize given patent images into pre-defined categories of images (such as graph, flowchart, drawing, etc.).

This document describes the last two tasks listed dealing with images.

2 Prior Art Image Task

This task is similar to the prior art search task, however the prior art patents should be found based on images and text of a query patent. As this is the first time this task is running, a relatively small set of patents have been chosen so as to enable the participants to gain first experience with patents and their images before being overwhelmed with data.

Three IPC sub-classes[1] are chosen for which patent searchers often rely on visual comparison of images in the patents to find relevant prior art. The IPC sub-classes included in the data are:

[1] Information on IPC classification can be found here: http://www.wipo.int/classifications/ipc/en/

A43B	CHARACTERISTIC FEATURES OF FOOTWEAR; PARTS OF FOOTWEAR
A61B	DIAGNOSIS; SURGERY; IDENTIFICATION
H01L	SEMICONDUCTOR DEVICES; ELECTRIC SOLID STATE DEVICES NOT OTHERWISE PROVIDED FOR

The target data set contains all images for patent documents in these three IPC sub-classes that have an application date previous to 2002 (291,566 images grouped by patent document). Note that each patent usually has more than one image. A patent is usually classified into more than one IPC subclass.

The queries will consist of the text and complete set of images of 211 patents. The aim will be to find the prior art for these patents based on both image and text information, or alternatively image information alone.

2.1 Target Data

The target data consists of European Patent Office (EPO) patents from the above three IPC classes that have an application date before 2002. Each patent is identified by a unique patent number. 6 archives must be downloaded to get all of the target data, 3 contain the images and 3 contain the associated patent xml files.

2.1.1 XML patent data

Corresponding to each patent is a directory containing the patent documents related to that patent. The layout is data/MAREC/EP/*nnnnnn*/*nn*/*nn*/*nn*/*.xml. Each patent can consist of a number of xml documents, giving the various states of the document during the application process. There are three archive files containing the XML documents, one for each IPC subclass (note that some patents occur in two archive files as a patent document usually has more than one IPC sub-class assigned to it).

For example, to patent EP 0983734 corresponds the directory containing files EP-0983734-A1.xml, and EP-0983734-B1.xml:

```
> pwd
EP/000000/98/37/34
> ls
EP-0983734-A1.xml EP-0983734-B1.xml
```

2.1.2 Image patent data

The images corresponding to the above patents are stored in three archives, also one for each IPC subclass. A patent usually has more than one image associated with it. The images corresponding to patent number EP*nnnnnnnnnnnn* (where the xml files are stored in the data/MAREC/EP/*nnnnnn*/*nn*/*nn*/*nn*/ directory)

have names EP*nnnnnnnnnnnn**.tif where the * is replaced by an alphabetic or numeric identification code for each image.

The images corresponding to the patent in the example above are the files `EP000000983734*.tif`.

About patent documents

In general, one patent (identified by a unique patent number) corresponds to several patent documents generated at different stages of the patent's life-cycle. Each stage is denoted by a kind code together with a version number[2]. The most common kind codes[3] are the following:

A1 publication of application with search report

A2 publication of application without search report

A3 publication of search report

A4 supplementary search report

A8 corrected title page of an EP–A document

A9 complete reprint of an EP–A document

B1 granted patent

B2 granted patent after modification

B8 corrected front page of an EP–B document

B9 complete reprint of an EP–B document

The "B" kind codes mark granted patents, and they necessarily have been earlier published as an "A1" or "A2" document.

2.2 Prior Art Image Task Description

The Prior Art Image Search Task (IMG_PAC) consists in finding documents in the target collection that may invalidate a given patent application. In intellectual property language: find documents that may constitute prior art for a topic patent. The topic document is a patent application document, A1 or A2. A PAC topic file contains a concatenation of topics. The structure of one topic is as follows:

[2] For the EP patents, documents at different stages have the same numeric identifier. For other patent offices this is not always the case. For example, the patent document US-6689545-B2 represents a US granted patent with its application document publication number US-2003011722-A1

[3] A full list of kind codes can be found at http://tinyurl.com/EPO-kindcodes

```
<topic>
<num>EP-1424100-A1</num>
<narr>Find all patents in the collection that potentially
invalidate patent application EP-1424100-A1.</narr>
<file>EP-1424100-A1.xml</file>
</topic>
```

where <num> contains the unique topic identifier consisting of the patent number, which itself contains a country code (always EP in this data set), a seven-digit number and the kind code (A1, A2). The <file> tag contains the name of the XML file.

The IMG_PAC_topics.xml file contains these topics. The patent topic xml files are found in the same directory as this xml file. The images corresponding to this topic patent are found in the sub-directory beginning with the patent number and ending with a date. For the example in the topic above, all images corresponding to patent EP-1424100-A1 are found in the directory EP-1424100-A1-20040602.

Note that the aim is to return the *patent documents* (identified by the patent number) corresponding to the topic, not the images themselves. It should be taken into account that a single patent document has more than one image.

The submission format is described in a separate document.

3 Image Classification Task

Images are an essential component of patents, as they illustrate key aspects of the invention. There are many different types of image in patents, including technical drawings, photos, flow charts, and graphs.

However, even though in many applications it is important to focus an analysis on a specific type of image, the annotation of the images according to the type in patents is in general either non-existant or poor with many errors.

The aim of the image classification task (IMG_CLS) is to automatically classify patent images according to type based on visual content. Manually classified and checked data is provided for training, and the long term aim is, based on these training data, to make it possible to reliably classify the millions of images in patents.

The classification is into 9 classes:

- abstract drawing
- graph
- flow chart
- gene sequence
- program listing
- symbol

- chemical structure
- table
- mathematics

Training data with between 300 and 6,000 training images for each of these classes is provided (see description below). Only these data may be used to train image type classification techniques.

At a later stage, we will publish a test database of 1,000 images. For each of these images, participating groups are required to determine the type of image.

3.1 Training data

The training data, organised into 9 directories - one for each class - contains the following number of training images per class:

Class	Class Number	Abbreviation	# Training Images
drawing	1	ad	5566
chemical structures	2	cf	5958
program listing	3	cp	5574
gene sequence (dna)	4	dn	5983
flow chart	5	ff	311
graph	6	gr	1664
mathematics	7	mf	5950
table	8	tb	5502
character (symbol)	9	tx	1579

3.2 Test data

Test data consists of 1,000 unclassified images. It has been released.

3.3 Format of the submission

This is available in a separate document.

3.4 Evaluation

Please note that it is not permitted to use any additional data for training and setup of the systems. If you need test data for system tuning, you need to split the available training data into a training and validation set. We will use equal error rate (EER) and Area Under Curve (AUC) of a ROC curve and True Positive Rate (TPR) per class averaged over all classes to evaluate the performance of the individual runs.