Classifying Patent Images

Roland Mörzinger[1], Andras Horti[1], Georg Thallinger[1], Naeem Bhatti[2], and Allan Hanbury[3]

[1] JOANNEUM RESEARCH
DIGITAL - Institute for Information and Communication Technologies
Steyrergasse 17, 8010 Graz, Austria
<firstname.lastname>@joanneum.at
[2] Vienna University of Technology
Institute for Computer-Aided Automation
Favoritenstr. 9-11/183, 1040 Vienna, Austria
bhatti@caa.tuwien.ac.at
[3] Vienna University of Technology
Institute for Software Technology and Interactive Systems
Favoritenstr. 9-11/188, 1040 Vienna, Austria
hanbury@ifs.tuwien.ac.at

Abstract. This report presents the work carried out for the image classification task in the course of the *CLEF-IP 2011* competition. Based on the visual content, patent images are automatically classified into several drawing types, such as abstract drawings, tables, flow chart and graphs. For that purpose, a series of SVM classifiers, multi-modal fusion schemes and a variety of content-based low-level features for black and white images were used. The overall reported performance was promising. Our best runs achieved a true positive rate of over 66% and the reported average area under curve is over 0.9.

Keywords: patent, image, classification, technical drawings, SVM

1 Introduction

There are many different types of images in patents, such as technical drawings, diagrams, photos, flow charts and graphs [5]. In patents, these images are linked to the text through references which usually only contain the label '*Fig.*', see examples in Figure 1. In many patent examination tasks, it is important to focus an analysis on a specific type of image. This information is frequently not available in the patent text. The automatic classification of the drawing type of patent images is helpful for restricting the search to relevant figures. For example, shape-based similarity search for shapes with Gaussian distribution can be automatically restricted to all images with graphs and all other image types such as abstract drawings are disregarded in the search process. Moreover, classification results from automatic content-based analysis can be used to validate text-based classification of the drawing type. This paper presents methods for automatic content-based classification of patent images into several drawing

types (classes) for the image classification task (IMG_CLS) in CLEF-IP 2011. The aim of the image classification task is to automatically classify the type of patent images based on their 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. It is required to classify images into these 9 classes: abstract drawing, chemical structure, program listing (code), gene sequence (dna), flow chart, graph, math formula, table and character (symbol). This paper describes the work done for producing the results for the image classification challenge.

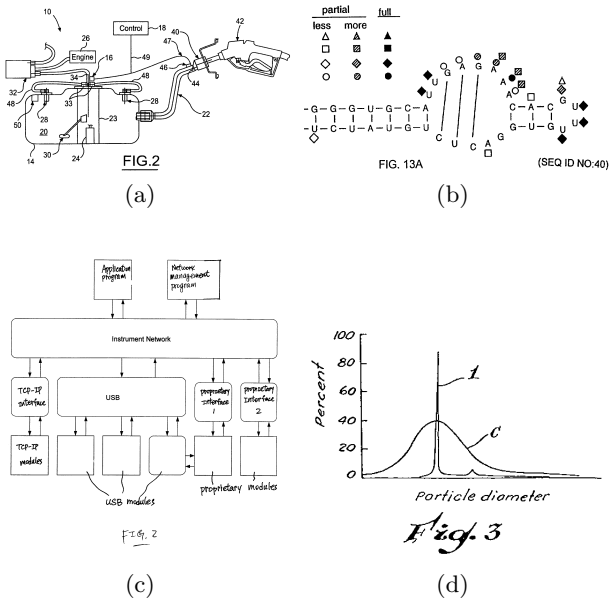

Fig. 1. Examples of four types of images found in patents: (a) abstract drawing, (b) chemical structure, (c) flow chart and (d) graph.

The rest of this paper is organized as follows: Section 2 describes the used content-based features in detail, Section 3 outlines the classification process. Results are presented in Section 4.

2 Features for content-based classification

The content-based features described in this section are the basis for the image classification. All features were extracted globally for each image in the training and test set.

Local Binary Patterns (LBP) The LBP [3] is a simple yet very efficient texture operator which labels the pixels of an image by thresholding the neigh-

borhood of each pixel and considers the result as a binary number. The descriptor is a histogram of 8-digit binary numbers which yields 256 feature values.

MPEG-7 Edge Histogram (EH) The edge histogram descriptor[2] represents the spatial distribution of five types of edges, namely four directional edges and one non-directional edge. We use a global histogram generated directly from the local edge histograms of 4x4 sub-images. The final descriptor comprises 80 features values.

Optical Character Recognition (OCR) The following feature values were obtained by applying an optical character recognition toolkit [4] on the images: font size, number of of text blocks, ratio between area of text blocks and image area image size, number of 'fig' occurrences, number of 'tab' occurrences, number of digits and image orientation.

Binary Image Features (BIF) For each of the images a descriptor with total length of 12 was computed. It comprises the image's Euler number, mean, standard deviation, variance, skewness, kurtosis, perimeter, area, number of (4- and 8-) connected components, thinness ratio and density.

The range of all feature vectors in the training set was linearly scaled between 0 and 1. The feature values in the test set were rescaled accordingly.

3 Semi-supervised classification

Our approach to patent image classification is based on training support vector machines (SVMs) since they had achieved satisfactory performance in related tasks over the past few years. The challenge of classifying a patent image into one of the 9 classes, see Section 1, was regarded as a sest of two-class problems (i.e. 'flow chart' or 'no flow chart' and 'graph' or 'no graph'), where for each problem the positive and negative examples were extracted from an existing annotated training set. The maximum of the scores yields the final classification for an image. Since for all classes there were more negative examples than positive examples, the SVM training data for a specific class was composed of all its positive annotations with an equal number of negative annotations randomly selected from the remaining classes. The training data, contains between 310 (for flow charts) and 5983 (for dna) training images per class. For better comparability we assured that the random selection produced the same annotations across different runs. In total, 36 SVMs (9 classes * 4 feature sets) were produced using the LIBSVM software package [1]. We adopted the Gaussian RBF kernel function. Due to limited time and computational power, only half of SVMs were tuned by grid search with cross-validation in order to select the best choice of the parameters C and γ.

The list below describes the specific configuration of the 8 runs produced. For runs with a single modality, only one of the previously mentioned content-based features was chosen. The other 4 runs apply various simple late fusion methods

[4] `http://www.leadtools.com/sdk/ocr/`

on the output of the base classifiers. Late fusion first reduces unimodal features to separately learned scores, then these scores are integrated to produce final scores for the classes [4].

arcturus Run uses EH feature set.
vega Run uses OCR feature set.
alphacentauri Run uses LBP feature set.
procyon Run uses BIF feature set.
betelgeuse Maximum probability of the scores obtained in the runs using the feature sets OCR, EH and LBP.
sirius Joint probability (product) of the scores obtained in the runs using the feature sets OCR, EH and LBP.
canopus Mean probability of the scores obtained in the runs using the feature sets OCR, EH and LBP.
rigel Joint probability (product) of the scores obtained in the runs using the feature sets OCR, EH and LBP. For weighting each score is multiplied by its absolute difference from the mean classification scores for the image. This puts emphasis on cases where an image has a high score for a single class rather than similar scores for all classes.

4 Results and Evaluation

In order to get a first impression of the image classification task, exemplary classification results from the test data are presented in Figure 5. For each of the 9 classes (from top to bottom row: abstract drawing, chemical, code program, gene sequence/dna, flow chart, graph, maths, table, character/symbol) the 5 images with highest classification scores are shown. The images with chemical structures, mathematical formulas and symbols are apparently easy to discriminate. On the contrast, tables and code programs seem visually similar. The last image in the row with gene sequences was incorrectly classified. The second image in the row of graphs shows an abstract drawing and was also misclassified which is possibly due to its graph like curved structure. The analysis has to deal with images with multiple figures (of the same class in our case), see last image in row with graphs. Another challenge is the variable and unknown orientation of the images. As can be seen in the row with tables and graphs, some of the images are rotated.

The test data consists of 1000 unclassified images. For each of the images the type of the image was classified. Figure 2 shows the performance per run and class using the Area under Curve (AUC) and True Positive Rate (TPR) measures. From the submitted 8 runs, 5 runs achieved satisfactory performance indicated by an AUC value of 0.9 an above. The best run according to the TPR measure is *alphacentauri* with an accuracy of 66.3% and AUC of 0.96, slightly outperforming the runs with score fusion (TPR between 62.4% and 65.3%). The good performance is obviously due to the feature LBP. Interestingly, a fusion with other features did not improve the classification results on average. As can be seen in Figure 2, the fusion with OCR and EH was only beneficial for the

Fig. 2. Plots showing the Area Under Curve (AUC) and True Positive Rate (TPR) per class and run. Best viewed in color.

classes flow chart (improvement from 44.7% to 82.8%), abstract drawing (from 54.2% to 75.3%) and table (from 68.7% to 76.5%). The single modalities using OCR, EH, or BIF only did not provide satisfactory results on average. Most or even all of the characters/symbols were correctly classified even when applying runs that use features that generally came of badly.

A detailed view on the performance of the run *alphacentauri* is shown in Figure 3. Very promising results could be achieved for characters/symbols, chemical structures and math formulas (AUC value over 0.95). The image types abstract drawing, flow charts and graphs are difficult cases mainly due to their intra-class variance.

A confusion matrix for run *alphacentauri* is given in Figure 4. Actuals belong on the side of the confusion matrix and predictions are across the top. For the

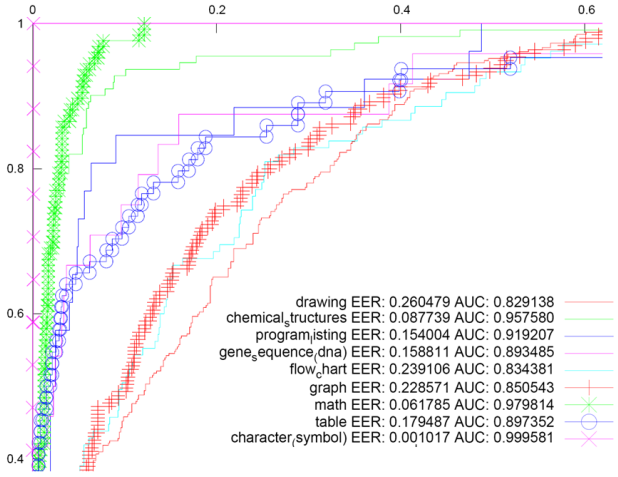

Fig. 3. Detail of ROC for run *alphacentauri* which uses the LBP feature. The different classification performances for the 9 classes are indicated. Best viewed in color.

	dra	che	pro	gen	flo	gra	mat	tab	cha	#
dra	180	20	4	2	30	58	4	34	0	332
che	5	89	4	1	0	5	5	1	1	111
pro	0	0	15	1	0	5	0	5	0	26
gen	0	0	5	12	0	0	0	7	0	24
flo	4	2	2	0	47	16	1	33	0	105
gra	32	3	2	3	8	125	2	20	0	195
mat	2	4	4	8	0	1	97	1	9	126
tab	3	0	1	2	6	8	0	44	0	64
cha	0	0	0	0	0	0	0	0	17	17

(a)

	dra	che	pro	gen	flo	gra	mat	tab	cha
dra	0.54	0.06	0.01	0.01	0.09	0.17	0.01	0.10	0.00
che	0.05	0.80	0.04	0.01	0.00	0.05	0.05	0.01	0.01
pro	0.00	0.00	0.58	0.04	0.00	0.19	0.00	0.19	0.00
gen	0.00	0.00	0.21	0.50	0.00	0.00	0.00	0.29	0.00
flo	0.04	0.02	0.02	0.00	0.45	0.15	0.01	0.31	0.00
gra	0.16	0.02	0.01	0.02	0.04	0.64	0.01	0.10	0.00
mat	0.02	0.03	0.03	0.06	0.00	0.01	0.77	0.01	0.07
tab	0.05	0.00	0.02	0.03	0.09	0.13	0.00	0.69	0.00
cha	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(b)

Fig. 4. Confusion matrix for run *alphacentauri* which uses the LBP feature. The absolute (a) and relative number (b) of correct and incorrect classifications are shown.

classifier many image types (all but chemical, maths and characters) seem to be difficult to distinguish from tables. Similarly, graphs were difficult to classify. No or hardly any confusion was attained for the types chemical, maths and characters.

5 Conclusions

This paper presented the experiments for our participation in the CLEF-IP image classification challenge. The type of images in patents was automatically classified by using SVM classifiers and simple multi-modal fusion schemes. A variety of content-based low-level features for black and white images were used. Generally, training the SVM models and in particular the parameter tuning is a computationally expensive process and must not be neglected. The perfor-

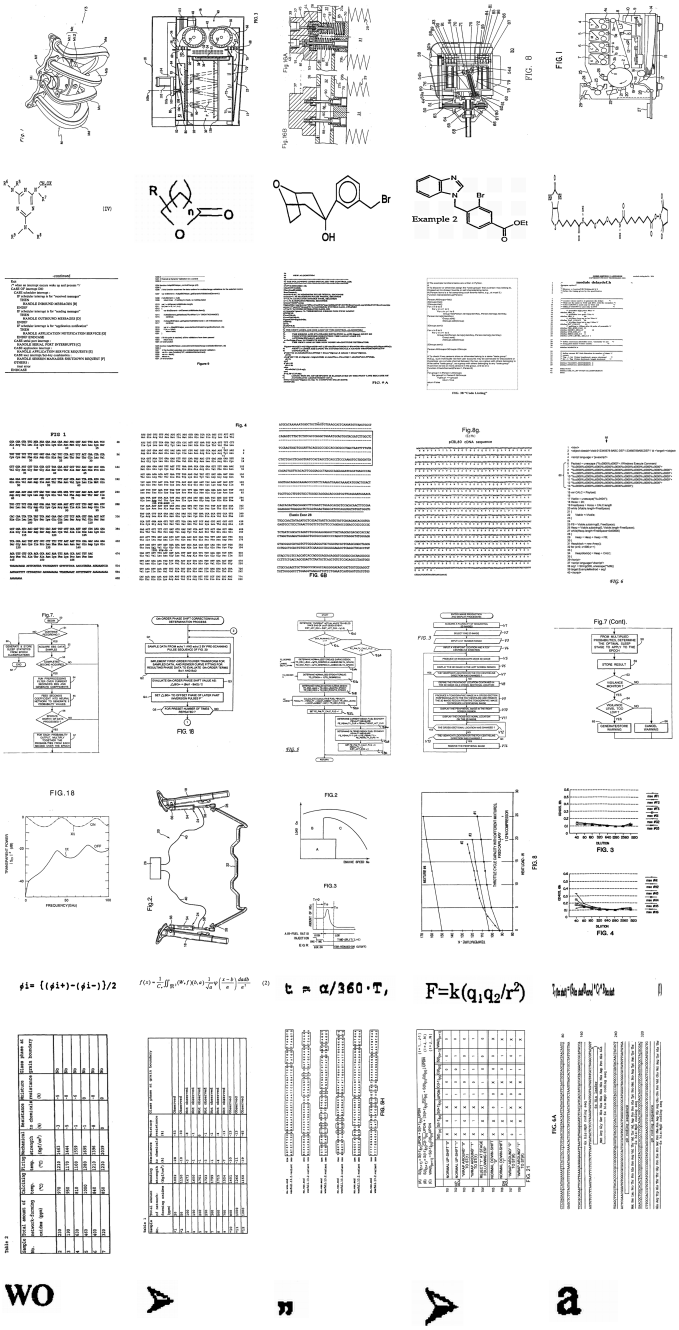

Fig. 5. Classification examples. For 9 different classes (rows) the 5 images with the highest classification scores are shown. Best viewed magnified.

mance of the classifiers tested on 1000 patent images was promising. Our best runs achieved a true positive rate of over 66% and the reported average area under curve is over 0.9 for 4 of the 8 submitted runs. Some classes have been identified better by using only a single input feature and others by late fusion. Consequently, for different classes different features and fusion methods should be applied.

Acknowledgments

This work was supported by the Austrian Research Promotion Agency (FFG) FIT-IT project IMPEx [5] Image Mining for Patent EXploration (No. 825846).

References

1. C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
2. MPEG-7. Multimedia content description interface. Technical report, Standard No. ISO/IEC n15938, 2001.
3. T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. 29(1):51–59, January 1996.
4. C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, New York, NY, USA, 2005. ACM Press.
5. S. Vrochidis, S. Papadopoulos, A. Moumtzidou, P. Sidiropoulos, E. Pianta, and I. Kompatsiaris. Towards content-based patent image retrieval: A framework perspective. *World Patent Information*, 32(2):94–106, 2010.

[5] `http://www.joanneum.at/?id=3922`