Concept Extraction from Patent Images Based on Recursive Hybrid Classification

Anastasia Moumtzidou, Stefanos Vrochidis, and Ioannis Kompatsiaris

CERTH-ITI, Thessaloniki, Greece
{moumtzid,stefanos,ikom}@iti.gr

Abstract. Recently, the intellectual property and information retrieval communities have shown interest in patent image analysis, which could augment the current practices of patent search by image classification and concept extraction. This article presents an approach for concept extraction from patent images, which relies upon recursive hybrid (text and visual-based) classification. To evaluate this approach, we selected a dataset from the footwear domain.

Keywords: patents, images, concepts, classification.

1 Introduction

The growing number of patent applications submitted worldwide necessitates the development of advanced patent search technologies. Recently, the Intellectual Property and the Information Retrieval communities have shown great interest in patent image search, expressed with common research activities in relevant conferences (e.g. IRFC[1], CLEF-IP[2]). Non-textual elements play a crucial role in patent search, since image examination is important to patent searchers to understand the contents and retrieve relevant patents. One of the first systems dealing with patent image search was PATSEEK [1], while more recently PatMedia [2] image search engine was developed. Following the recent challenges in image analysis (semantic indexing, semantic gap), the latest approaches in patent image search deal with patent image classification [3] and concept extraction [4]. However, the motivation behind the interest in patent concept-based search is also revealed by the following scenario [5]: a patent searcher searches for a dancing shoe that incorporates a rotating heel with ball bearings; at first, the patent searcher recognises the main concepts of the invention (e.g. dancing shoe) and based on them defines keywords and relevant classification areas. In many cases the important information is described with figures. Therefore, it would be important if the patent searcher could directly retrieve patents, which include figures depicting these concepts. Such concept-based retrieval functionalities could be integrated in existing patent search systems to facilitate the tasks of patent searchers.

[1] Information Retrieval Facility Conference.
(http://www.ir-facility.org/irf-conference-2012).
[2] http://www.ir-facility.org/clef-ip

M. Lupu, E. Kanoulas, and F. Loizides (Eds.): IRFC 2013, LNCS 8201, pp. 83–86, 2013.

In this paper, we present an approach for concept extraction from patent images based on a supervised machine learning framework using Support Vector Machines (SVM) trained with textual and visual features. This work goes beyond [4] by proposing a recursive scheme for concept extraction and deals with a more complicated scenario compared the classification task of CLEF-IP 2011 (e.g. [3]), in which more generic and visually dissimilar categories (e.g. flowcharts, symbols) are considered.

2 Patent Image Concept Extraction Framework

The proposed framework (testing phase) is depicted in Figure 1. The initial step includes the extraction of all patent images and the associated captions. Then, the images are fed into the feature extraction component, where visual and textual features are generated. Subsequently, the dataset is annotated and separated into training and test set. During this step the images of the same patents are kept together. Then, three models are trained for each concept using: a) visual features, b) textual features, and c) the results of the previous models. The latter are used as features to drive a hybrid classification model to provide the final results. The test (validation) set is further split into two sets (i.e. A and B) based on the following rule: the figures with description that points to another figure of same patent (e.g. "*Fig. 2* is the front view of *Fig. 1*") belong to set B, whereas the rest to set A. This kind of descriptions is one of the main reasons for retrieving low quality results in text-based concept extraction [4]. During the testing phase (Figure 1), the figures of set A are fed into the textual and visual feature extraction components and their features are used as input to the textual and visual classifiers respectively. The final confidence score for each concept is provided by the hybrid classifier. Then, the missing parts of the descriptions of the images contained in set B are replaced by the results of the textual classifier in a recursive way. For instance, in the previous example, if "*Fig. 1*" is annotated as "ski boot", the new caption of "*Fig 2*" will be: "*Fig. 2* is the front view of the of ski boot". This process continues recursively until all the captions of the figures in set B are updated. Finally, the figures of set B are processed in a similar way with the ones of set A.

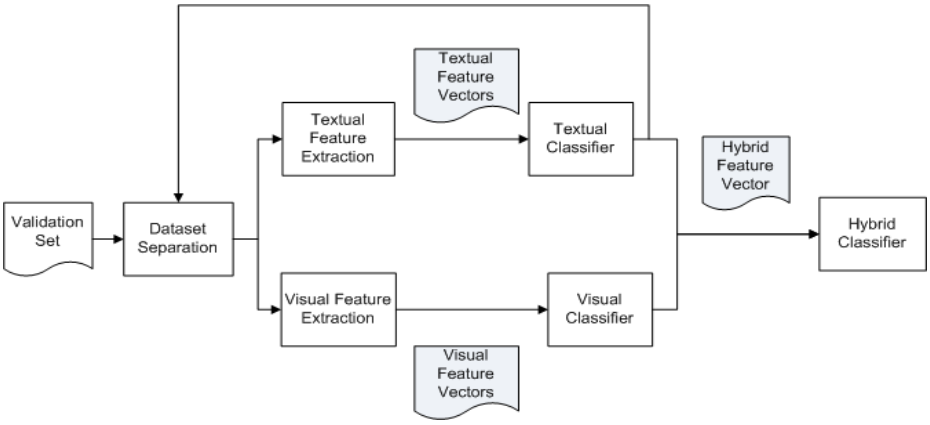

Fig. 1. Testing procedure of patent concepts

The LIBSVM [6] library was used for SVM classification. The global visual features employed are the Adaptive Hierarchical Density Histograms, which have shown discriminative power for patent images ([2], [4]). The textual feature extraction is based on the bag-of-words approach. We produce a 400-term lexicon, which includes the most frequently used words of this dataset. Indexing is performed using Lemur[3]. The weight $w_{t,d}$ of each term t for figure caption d is calculated as follows:

$$w_{t,d} = \frac{frequency\ of\ term\ t\ in\ caption\ d}{total\ number\ of\ words\ in\ caption\ d} \quad (1)$$

3 Results and Evaluation

To evaluate the approach, we use an annotated dataset (described in [4]) extracted from 355 patents. It contains around 1000 patent images depicting parts of footwear. Each image is associated with a single concept. The concepts selected are shown in Table 1, and represent specific International Patent Classification (IPC) groups (A43B21, A43B23, A43B5). Examples for ski boot concept are provided in Figure 2.

We apply three-fold cross validation at the patent level. Each training set consists of around 650 and the testing of around 250 images (in average 200 from set A and 50 from set B). Then, we evaluate the results by presenting the accuracy and F-score of the concept detectors. Table 1 contains these metrics for the validation set (A∪B). The results show that the hybrid model demonstrates better performance in terms of accuracy and F-score, while the textual features report higher F-score than the visual ones.

Table 1. Fscore and Accuracy for the validation set (A∪B)

Concept	Visual		Textual		Hybrid	
	F-score	**Accuracy**	**F-score**	**Accuracy**	**F-score**	**Accuracy**
Cleat	42,63%	88,64%	57,92%	90,58%	58,40%	89,02%
Ski boot	63,73%	92,34%	78,91%	95,55%	77,22%	95,59%
High heel	55,73%	90,50%	54,47%	89,05%	66,01%	91,96%
Heel with spring	50,63%	93,04%	51,50%	92,59%	53,24%	93,07%
Toe caps	43,66%	90,53%	74,59%	95,37%	72,45%	94,59%
Average	**51,28%**	**91,01%**	**63,48%**	**92,63%**	**65,46%**	**92,85%**

With a view to evaluating the performance of the recursive approach followed for set B, we compare the results of the proposed approach with the baseline (similar to [4]). Table 2 contains the results of both approaches, which shows that the recursive classification-based approach outperforms the baseline. Finally, Figure 2 depicts the first six figures with higher prediction scores for the concept "ski boot".

[3] Lemur Project, `http://www.lemurproject.org/`

Table 2. Performance measures F-score and accuracy for validation set B

Concept	Baseline approach		Proposed recursive approach	
	F-score	Accuracy	F-score	Accuracy
Cleat	59,95%	85,42%	77,23%	91,15%
Ski boot	57,78%	95,51%	57,78%	95,51%
High heel	71,64%	91,50%	76,23%	93,89%
Heel with spring	53,74%	91,71%	53,74%	91,71%
Toe caps	66,67%	93,46%	85,19%	97,69%
Average	61,96%	91,52%	70,03%	93,99%

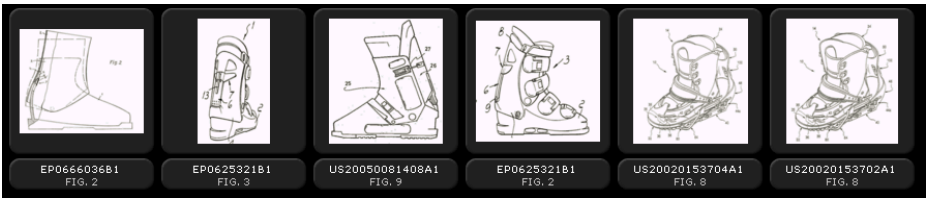

Fig. 2. Results for concept "ski boot"

4 Conclusions

In this paper, we have presented an approach for concept extraction from patent images based on recursive classification. The concept retrieval module could be a part of existing patent search systems, in order to support patent searchers in patent invalidation and valuation tasks. Future work includes testing the method by considering a larger patent database and additional concepts, in order to further test its scalability.

References

1. Tiwari, A., Bansal, V.: PATSEEK: Content Based Image Retrieval System for Patent Database. In: Proc. International Conference on Electronic Business, Beijing, China (2004)
2. Vrochidis, S., Papadopoulos, S., Moumtzidou, A., Sidiropoulos, P., Pianta, E., Kompatsiaris, I.: Towards Content-based Patent Image Retrieval; A Framework Perspective. World Patent Information Journal 32(2), 94–106 (2010)
3. Mörzinger, R., Horti, A., Thallinger, G., Bhatti, N., Hanbury, A.: Classifying patent images. In: Proceedings of CLEF 2011, Amsterdam (2011)
4. Vrochidis, S., Moumtzidou, A., Kompatsiaris, I.: Concept-based Patent Image Retrieval. World Patent Information Journal 34(4), 292–303 (2012)
5. De Marco, D.: Mechanical patent searching: a moving target. In: Patent Information Users Group (PIUG), Baltimore, USA (2010)
6. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011)