

Concept-based patent image retrieval

Stefanos Vrochidis*, Anastasia Moumtzidou, Ioannis Kompatsiaris

Informatics and Telematics Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece

A B S T R A C T

Keywords:

Concepts
Visual
Textual
Patent
Classification
Content-based search
Retrieval
Drawings
Image
Search engine
Hybrid

Recently, the intellectual property and information retrieval communities have shown increasing interest in patent image retrieval, which could further enhance the current practices of patent search. In this context, this article presents an approach for automatically extracting concept information describing the patent image content to support searchers during patent retrieval tasks. The proposed approach is based on a supervised machine learning framework, which relies upon image and text analysis techniques. Specifically, we extract textual and visual low-level features from patent images and train detectors, which are capable of identifying global concepts in patent figures. To evaluate this approach we have selected a dataset from the footwear domain and trained the concept detectors with different feature combinations. The results of the experiments show that the combination of textual and visual information of patent images demonstrates the best performance outperforming both single visual and textual features results. The outcome of this experiment provides a first evidence that concept detection can be applied in the domain of patent image retrieval and could be integrated in existing real world applications to support patent searching.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Nowadays, the growing number of patent applications submitted in patent offices worldwide calls for the need of advanced patent search technologies, which could deal with the complexity and the unique characteristics of patents. Despite the fact that the patent documents contain multimodal information encoded in textual, tabular and figure format, most of the developed techniques and patent search engines to date, still rely only upon text to provide retrieval functionalities. To a certain extent, indeed textual data can be considered as a very reliable source of information, since the ideas and the innovations to be patented are almost always described in such a format in the claims and the disclosure parts of the patent. However, most of the patents include a drawings section, which contains figures, drawings and diagrams as a means to further describe innovative artifacts, processes, algorithms and other inventions. Recently, both the intellectual property (IP) and the information retrieval (IR) communities have shown great interest in patent image search expressed with research activities and works in the area (e.g. [1,2]), as well as with prototype systems and demos (e.g. [3,4]). In addition, dedicated

sessions and talks have been organised in relevant symposiums, workshops and conferences (e.g. IRFS,¹ CLEF,² etc.).

Non-textual elements of patents can play a crucial role in patent search [5]. Specifically, image examination can be considered very important to patent searchers in their attempt to understand the patent contents and retrieve relevant patents. During this procedure there are several cases, in which patent searchers are browsing thousands of patents looking only on the images contained in the drawings section. Such tasks could be addressed and speeded up with the aid of patent image search engines, which could retrieve images based on their visual content. An additional reason that patent image search could be of great importance, is the fact that images by nature are independent of the applicant's language and remain intact despite the evolution of the scientific terminology over the years. This means that patent searchers could be able to detect relevant multilingual patents without needing a translation, which in many cases is either difficult and time consuming to be generated (e.g. in Asian patents) or the automatic translation does not yield quality results. Despite the recent advances of machine translation (MT) and the recent cooperations (e.g. agreement between EPO and Google) towards the application

* Corresponding author. Tel.: +30 2311 257754; fax: +30 2311 257707.
E-mail address: stefanos@iti.gr (S. Vrochidis).

¹ Information Retrieval Facility Symposium (IRFS).

² Conference on Multilingual and Multimodal Information Access Evaluation (CLEF).

of automatic patent translation, the latter still has its limitations. Specifically, MT could reveal the gist of the invention, but in many cases it could not cope with the variant terminology and expressions, which might depend on context and word sequence in order to provide a reliable and accurate result. In addition, patent image based search can support more effectively the comparison and retrieval of patents published in different chronological periods (e.g. decades), while traditional text-based search might fail due to the variant terminologies changing over time.

Motivated by this situation, our previous work [3] was directed towards the research and the development of a system, which could automatically extract and retrieve patent images based on visual similarity. Although the functionality of retrieving similar images was considered very useful by professional patent searchers [6], there are many situations, in which the actual need is to identify images with common characteristics that fall into a specific category or illustrate a specific object or concept. For instance, in many cases a patent searcher is submitting textual queries looking for figures that depict specific schemas or objects (e.g. high heel shoes, sky boots, etc.). In order to fully address this requirement we need to understand what exactly a certain figure depicts, not only based on the associated caption (if it is available) but also based on its content.

To this end, and following the trend of modern image retrieval approaches, which are moving towards concept-based image search [7], we investigate concept extraction in patent images. Concept-based patent image retrieval consists of several challenges. The most important of them are listed below:

1. Establishment of a taxonomy (i.e. organised set/hierarchy of concepts) for the broad patent domain
2. Extraction of descriptive features from patent images that would minimise the effect of the semantic gap (i.e. the difficulty in translating low level features into human understandable concepts)
3. Addressing the problem of training set development required by supervised classification frameworks

The establishment of a concept taxonomy for the patent domain (first challenge) could be supported by existing patent-relevant resources including the IPC classification schema, patent ontologies [8], or even application of ontology learning on patent document corpuses. In addition, adaptation of existing image taxonomies (e.g. ImageNet³) to the patent domain could assist this procedure, although taxonomies of broad interests cannot capture very technical or scientific concepts (e.g. piezoelectric film). The proposed approach deals with the second challenge and applies image analysis and machine learning algorithms in the patent domain considering a predefined set of concepts. Specifically, we extract low-level textual and visual features (developed specifically to deal with complex drawings) from patent images and then we employ a supervised machine learning framework realised with Support Vector Machines (SVM). Then, we use manually annotated data to train a detector for each concept and finally we evaluate the approach using a dataset from the footwear domain. This approach requires the development of training sets (third challenge), which constitutes an important issue due to the manual effort required in such a process. However, it should be noted that the training set development is performed only once for each specific concept. Then they can be applied to large patent collections, however, still in several cases, new definition and re-training might be required. On top of that, recent approaches towards the solution of this

problem propose collaborative annotation initiatives and automatic development of training sets by exploiting surrounding text (e.g. captions) and user implicit feedback (e.g. mouse clicks and keyboard inputs) during search (e.g. [9]). However, the automatic development of training sets is beyond the scope of this article.

The main contribution and the research objective of this paper is the investigation of global concept extraction from patent images combining visual and textual data, given the complexity and the specific nature of patent images. A notable scientific contribution of this work is the application of the recently developed AHDH feature [10] in a SVM classification framework, in order to evaluate its performance in supervised problems.

The article is organised as follows. First, we present the related work in Section 2, while in Section 3 we discuss the use cases and the requirements involved in the task of concept extraction for patent images. The architecture of the proposed framework and the involved techniques are analysed in Section 4. Section 5 presents the experiments we have conducted, while the results are reported in Section 6. Finally, Section 7 concludes the paper.

2. Related work

To the best of the authors' knowledge, up to now no other attempt has been made for extraction of semantic concepts from patent images combining visual and textual information. The existing approaches in patent image search domain are dealing mostly with visual similarity retrieval and figure classification in broad categories. Most of these works have been recently proposed, while before 2007, no systematic efforts have been conducted with the aim of developing patent image retrieval systems [11]. Therefore, in this section we will discuss both the concept-based image retrieval in other domains, as well as the most relevant patent image search and classification approaches to date.

One of the grand challenges in multimedia information retrieval is the automatic visual concept detection. Concept-based multimedia retrieval is generally performed by exploiting information contained in heterogeneous sources. The two main approaches followed for concept generation and search in multimedia content are: "content-based" and "text-based". The content-based retrieval uses the analysis of low-level features [12] to represent the multimedia content. Such features are usually based on colour distribution, texture, edge, geometry etc. to describe the visual content. The indexing structure of low-level features allows retrieval employing the "query by example" methodology (i.e. finding pictures of an object based on an example photo of that object). More complex approaches include segmentation of the example image so the user can query a system by using as input only a specific region (i.e. the object of interest) [13]. Other related works (e.g. [14,15]) employed machine learning approaches, which train models with annotated development data (i.e. multimedia content manually labelled with specific concepts), in order to provide confidence coefficients regarding the existence of a concept. On the other hand, text-based retrieval uses the indexing of media according to text that can be associated to it, such as titles or descriptions in associated metadata files, or text found close to the media on a Web page. Such retrieval methods are also exploiting external resources to define concepts with the usage of synonyms, hyponyms and textual hierarchies. Although text-based retrieval can be considered as more reliable, it is highly based on the existence and the quality of the annotations, which can be noisy (e.g. tags from different users, etc.), or they do not describe adequately the visual content. Other recent works [16] combine the aforementioned techniques by employing fusion methods of the heterogeneous information and promising results have been presented.

³ ImageNet: <http://www.image-net.org/>.

The first patent search and classification systems were focussing on text- and metadata-based search. Only in the recent years there have been attempts towards patent image search. However, most of the image retrieval work in this area is dedicated to the field of trademark search (e.g. [17,18]). One of the first systems that attempted to tackle the patent image retrieval problem was PAT-SEEK [4], which is an image-based retrieval system for the US patent database applying a shape-based image retrieval method called the Edge Orientation Autocorrelogram (EOAC). Another related effort comes from a French company, LTU Technologies [19], which was used by the French patent office (INPI) to build an image-based patent retrieval system and was also applied in the eMARKS [20] project. More recently, the PatMedia [21] image search engine was developed during the PATExpert project [22]. PatMedia is capable of retrieving patent images based on visual similarity using the Adaptive Hierarchical Density Histograms [10] and constitutes the retrieval engine of an integrated patent image extraction and retrieval framework [3]. Finally, recent works have dealt with classification of patent images under generic categories (i.e. flowcharts, diagrams, etc.). Specifically the authors in [23] employed a semi-supervised classification approach based on Support Vector Machines and extracted many different image features including Local binary patterns, MPEG-7 Edge histograms, binary image features and image characteristics retrieved with Optical Character Recognition (OCR). In another approach [24], the authors extract local orientation histograms using variations of the Scale-Invariant Feature Transform (SIFT) algorithm and they build visual vocabularies specific to patent images using Gaussian mixture model (GMM). Then the images are represented by Fisher features and linear classifiers are employed for the categorisation.

The recent works in patent image classification and retrieval show that the IP and IR domains are mature enough to move to concept based patent image search.

3. Use cases and processing requirements

Before presenting the approach for concept based patent image search, it is essential to discuss the patent search practices to investigate how this new functionality could serve the needs of patent searchers. To this end we present use cases of patent search, which could benefit from concept-based retrieval and analyse the requirements that arise.

3.1. Patent search scenario

Patent searchers are experts at searching but not always regarding all the technologies and the areas in which they work. In this context the patent searchers need to learn the gist of an invention, the new and the old terminology, the multiple classifications and the parallel technologies, which constitutes a really difficult task. Let's present an example of a mechanical search as this is described by a professional patent searcher [25]. We assume that we have a disclosure:

"A dancing shoe with a rotatable heel to allow rapid pivoting about your heel. In a preferred embodiment, the heel should have ball bearings."

A patent searcher has to distil the gist for this disclosure, which could be the basis of the upcoming search. In this case the gist could be expressed by the following concepts:

Concept 1: Dancing shoe
 Concept 2: Rotating heel
 Refined Concept 2: Rotating heel with ball bearings.

Then, the patent searcher proceeds by defining specific keywords based on the aforementioned concepts and classification areas to search. In many cases, including this example, the important information (and the gist) are usually illustrated and described with the aid of figures. It is evident that it would be of great help if the patent searcher could directly retrieve patents, which include figures depicting these concepts (i.e. dancing shoe and rotating heel for this example).

3.2. Requirements

To support concept extraction from images we need to exploit the heterogeneous information provided in the patent for a specific figure. That is the image itself and the figure description. Although one could argue that the image description would be adequate for the concept extraction, this is not always the case due to several reasons. First, many figures can be associated with misleading or incomplete descriptions (e.g. references to other figures or parts of the patent). In addition, there are cases, in which it is not trivial to automatically map the figure caption to the corresponding image due to handwritten figure labels that cannot be automatically recognised. Finally, many patents and therefore the image descriptions are written in a language that cannot be easily translated (e.g. Chinese). In such cases we need to rely solely on image information. However, when the figure caption is available we can still process it to gain additional information about the image.

Concept and object retrieval is currently one of the most active topics in multimedia analysis and indexing research domain. The approaches usually employed are based on manually annotated examples, which could drive a supervised machine learning approach and support the system in learning the concepts. Depending on the dataset and the way a concept is depicted, one can distinguish between local concepts/objects and global concepts which describe a figure as a whole. To give an example we can imagine an electrical circuit, which includes capacitors, resistors, etc. In this case a capacitor can be considered as a local concept/object, while the circuit is the global concept. The extraction of local concepts or objects would require either prior segmentation of the image, which however doesn't ensure that the segmented areas would reflect to meaningful objects, or the extraction of local visual features and salient points, which require computationally expensive salient point mapping algorithms (e.g. in a typical retrieval case, K-Nearest Neighbour (KNN) [26] algorithm is used to find the point matches, and Random Sample Consensus (RANSAC) [27] to reject inconsistent matches). On the other hand, extraction of global concepts is based on global visual features, which describe the image as a whole. Based on what is actually required by the patent searcher both types of information would be important. In this work we will focus on the extraction of global concepts given the fact that the first step is to get the global understanding and the domain that a figure belongs to, and we plan to exploit this knowledge in a future work to extract local concepts or objects. For instance, if we obtain the knowledge that a figure depicts a circuit, it would be much easier to identify that a local object or concept illustrated in the figure is a capacitor and not a water filter.

From the functional point of view, a set of requirements with regards to the system performance has to be defined. First, the system should be scalable as it has to cope with vast amounts of content (in the order of millions of patent images). This means that the processing techniques have to be fast and efficient and therefore large feature representation vectors and computationally expensive fusion algorithms should be avoided. However, the representations of the patent images should be

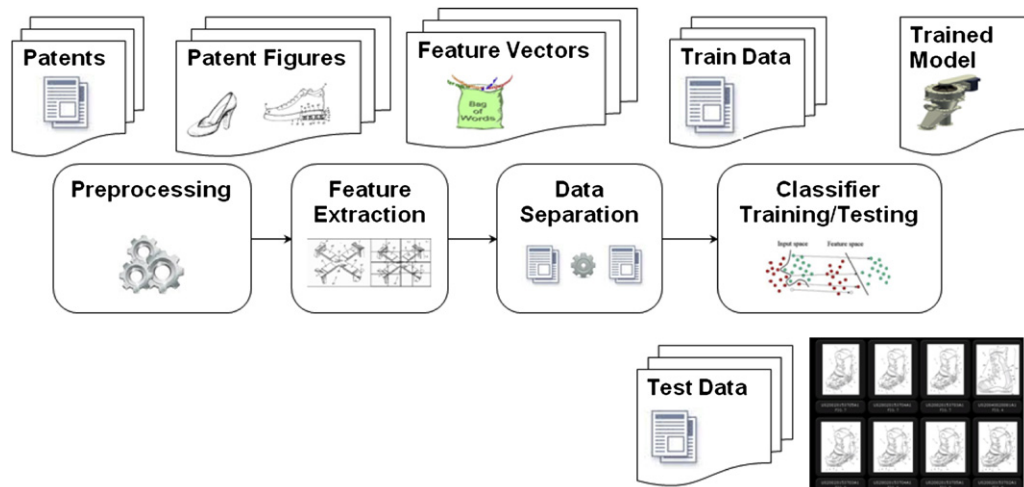


Fig. 1. Concept extraction framework. The chain of the components (e.g. Preprocessing) is illustrated in the middle, while the input and output (e.g. patents, feature vectors) are depicted below and above the chain.

adequate so that the concept detectors can demonstrate a minimum accuracy of 85–90% (depending also on the concept characteristics). In addition, a vast number of concepts that are characteristic for the patent images of each IPC class and subclass have to be defined by patent experts, while relevant examples have to be annotated to drive the machine learning algorithms. Finally, the framework needs to build upon open technologies and standards, in order to be easily adaptable to the established patent search platforms.

4. Patent image concept extraction framework

In order to meet the requirements described above, we propose a supervised machine learning-based framework that combines well established and state of the art techniques from text and image analysis. The proposed architecture is illustrated in Fig. 1.

The initial step includes document processing for extracting all the required images embedded in the document and related metadata. Although this could be performed with automatic segmentation techniques, which could decompose each document page of the drawings section into the figures it consists of (such segmentation techniques have been already applied in [3]), in this case we decided to apply a manual segmentation in order to test the concept detectors with quality data. Then the images are fed into the feature extraction component, where the visual and textual based features are generated. Subsequently, the dataset is manually annotated and separated into training and test set.

Finally, we train a classifier for each concept using the train data and we evaluate its performance using the test data. In the following we will describe in more detail the visual and textual feature extraction, as well as the supervised classification method involved.

4.1. Visual features

As we have already discussed, the extraction of global concepts requires the employment of global image features, which can deal with the complexity and the special characteristics of patent images. The main characteristic of patent figures is that they are mostly black and white and they depict technical information in diagrammatic form. Specifically, in the previous years the patent figures have been strictly black and white and only recently some Korean patent publication started to include colour images. However, even in this case, the colourisation does not follow specific standards and therefore the application of colour to gray-scale filters is required to address such figures, since the colour variation will not convey standardised information. Given the fact that general case image representation features are based on colour and texture, which are absent in most of patent images, we need to apply an algorithm which takes into account the geometry and the pixel distribution of these images. To this end, we apply the algorithm proposed in [10] to extract the Adaptive Hierarchical Density Histograms (AHDH) as visual feature vectors.

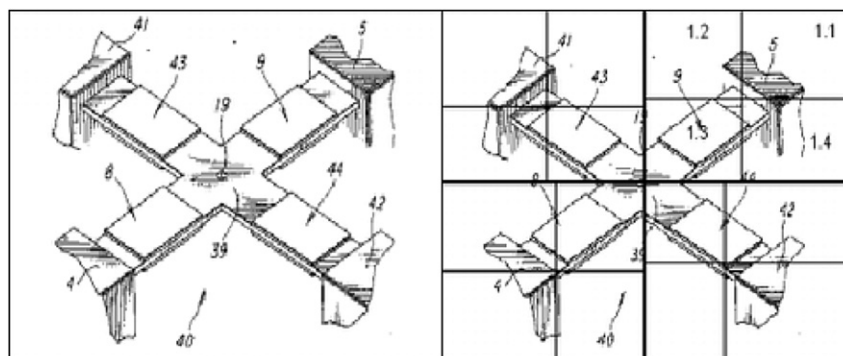
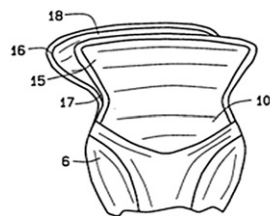


Fig. 2. Extraction of ADHD. Geometric centroids are utilised to iteratively split the image to new regions.



Example: Patent US 20020152637 A1

FIG. 7 shows the reversible tongue containing a pocket in its upper half, and which may be secured by Velcro, or the like, into closure

Fig. 3. Patent figure with the associated description.

The Adaptive Hierarchical Density Histograms (ADHD) were devised specifically to deal with such binary and complex images. The feature vector is generated based on the following steps. First, the algorithm involves a pre-processing phase for noise reduction, coordinates calculation and normalisation. After the pre-processing has taken place, the first geometric centroid of the image plane is calculated and the image area is split into four regions based on the position of this centroid. Subsequently, the feature vector is initialised by estimating the distribution of the black points in each region. This procedure is repeated in a recursive way (Fig. 2) for a manually specified number of iterations and after each iteration the feature vector is updated. This non-segmentation point-density orientated technique seems to combine high accuracy at low computational cost as it represents the image with a low dimension feature vector (i.e. around 100 features). Based on experiments conducted with patent datasets, the ADHD outperformed the other state of the art methods [10].

4.2. Textual features

In order to exploit the textual descriptions provided for each figure in the patent document, we process the captions and extract textual features for the patent images. Specifically, we apply a bag of words approach to model each figure with a vector. The bag of words model is a simplifying assumption used in natural language processing and information retrieval. In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order.

To generate such a vector we need to define a lexicon, which includes the most frequently used words of this dataset. Then for each figure and based on the associated description we calculate a weight for each word included in the lexicon. The textual annotations are processed with the aid of Porter stemmer [28] and the frequent stop words (e.g. and, so, etc.) are removed. The indexing of the remaining keywords is performed using Lemur [29]. The weight of each term is calculated with the well established metric tf-idf (term frequency multiplied with the inverse document frequency).

Assuming that the lexicon has the following format:

<boot snowboard illustr tongu footwear heel...>

then, the corresponding feature vector for the patent image illustrated in Fig. 3 would be:

[0 0 0 0.0909091 0 0...]

We can notice that only the 4th feature which corresponds to the stemmed (i.e. the keyword is reduced to its root by removing affixes) keyword “tongue” has a weight greater than zero as this is the only one that appears in the description.

In this approach we have selected a lexicon of 100 terms, which leads to the generation of feature vectors of the same size for all the patent images. The selection of the features was performed by thresholding the calculated weights of the words extracted from the dataset with experimental values.

4.3. Support Vector Machines

Support Vector Machines (SVMs) constitute a set of supervised learning methods, which analyse data and recognise patterns, and are employed to solve classification and regression problems. SVMs are a relatively new learning process that is influenced highly by advances in statistical learning theory and a sufficient increase in computer processing power in the recent years.

SVMs are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships (Fig. 4). Support Vector Machine (SVM) is primarily a method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels [30].

A schematic example is shown in Fig. 4. In this example, the objects belong either to class green stars or red circles. The separating line defines a boundary on the right side of which all objects are green and to the left of which all objects are red. Any new object falling to the right is labelled, i.e. classified, as green (or classified as red should it fall to the left of the separating line).

In this implementation, we make use of the LIBSVM library [31] and we consider a binary C-Support Vector Classification using as kernel the radial basis function.

5. Experiments

In order to investigate the potential of extracting concepts from patent images we tested the framework described in the previous section using a dataset from A43B and A63C IPC subclasses and we conducted several experiments.

5.1. Dataset and selected concepts

The dataset was manually extracted from around 300 patents, which contain parts of footwear. Based on the advice of professional patent searchers (acknowledged below) in this domain we have selected the following 8 concepts for this domain: cleat, ski boot, high heel, lacing closure, heel with spring, tongue, toe caps and roller skates, which represent specific IPC groups or subgroups. In Table 1 we present a more detailed description and a visual example of each concept.

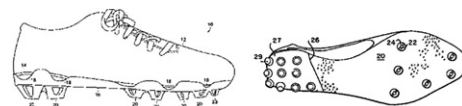


Fig. 4. Decision plane which separates the red circles class from the green star class. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Concepts description.

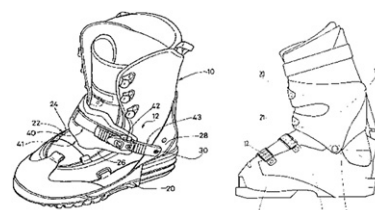
Cleat

A short piece of rubber, metal etc attached to the bottom of a sports shoe used mainly for preventing someone from slipping
IPC subgroup: A43B5/18S



Ski boot

A specially made boot that fastens onto a ski
IPC subgroup: A43B5/04



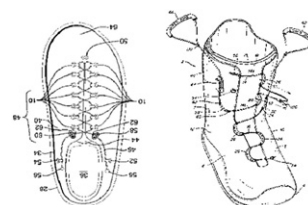
High heel

Shoes with high heels
IPC group: A43B21



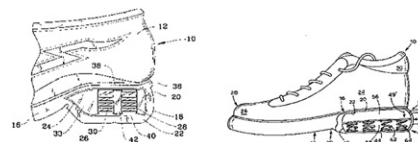
Lacing closure

A cord that is drawn through eyelets or around hooks in order to draw together the two edges of a shoe
IPC subgroup: A43B5/04



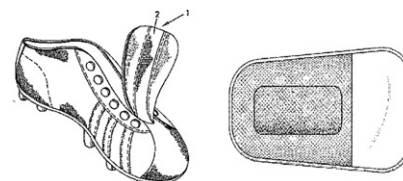
Spring heel

Heels with metal springs
IPC subgroup: A43B21/30



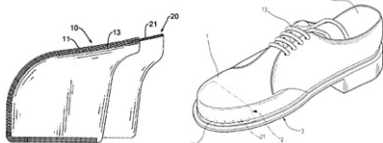
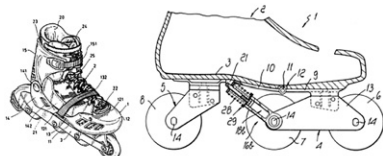
Tongue

The part of a shoe that lies on top of your foot, under the part where you tie it
IPC subgroup: A43B23/26



(continued on next page)

Table 1 (continued)

Toe caps
A reinforced covering of leather or metal for the toe of a shoe or boot
IPC groups: A43B23 and A43B7

Roller skate
A shoe or boot with two or four wheels or casters attached to its sole for skating on hard surfaces
IPC groups: A43B5 and A63C17


The segmentation of patent images and the association with the figure descriptions in the text have been done manually in order to have quality data to draw safer conclusions on the concept extraction method. Then, the images were manually annotated with the support and advice of professional patent searchers. In case a figure can be described by two or more concepts (e.g. a figure that depicts the “lacing” system on a “ski boot”), the assignment is given with respect to the purpose that the specific figure serves (e.g. whether the aim to describe the “lacing” system or the “ski-boot”). The numerical statistics of the dataset⁴ can be found in the Table 2.

5.2. Building the concept detectors

In this experiment we train one classifier for each concept. With a view to evaluating the performance of the textual and visual data in the concept extraction process, we have experimented with 3 different training approaches to build the concept detectors.

- Visual case: the classifier was trained only with visual features (Section 4.1)
- Textual case: the classifier was trained only with textual features (Section 4.2)
- Hybrid case: the classifier was trained with a hybrid feature vector, which was constructed based on early fusion approach. Specifically, the descriptor employed in this case was the result of the concatenation of the textual and visual feature vectors.

In order to optimise the training process and to select the appropriate parameters that could maximise the performance of the concept detectors, we have conducted a cross-validation process using the training set. In detail, the training set was divided into several (i.e. n) subsets of equal size and sequentially each subset is tested using the classifier trained on the remaining (i.e. $n - 1$) subsets. Then, the parameters that reported the best performance were selected for training. In this experiment the cross-validation was optimised to maximise the F -score metric (Section 6.1).

6. Results and evaluation

The presentation of the results and the evaluation procedure took place using the PatMedia search engine. In this section, we introduce the evaluation metrics we considered and we

Table 2
Dataset statistics.

Concepts	Total figures	Train figures	Test figures
Cleat	148	89	59
Ski boot	123	74	49
High heel	148	89	59
Lacing	117	71	46
Spring	106	64	42
Tongue	124	75	49
Toe caps	108	65	43
Roller	168	101	67
Total	1042	628	414

demonstrate quantitative evaluation in terms of performance metrics and visual results through user interaction modes.

6.1. Evaluation metrics

To evaluate the performance of the proposed approach, we analyse the results by presenting the accuracy of the concept detectors, the precision and recall of the results and the F -Score.

The accuracy of the concept detectors is calculated as:

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

where:

TP = True Positives, i.e. the relevant images that have been classified as relevant.

TN = True Negatives, i.e. the non-relevant images that have been classified as non-relevant.

FP = False Positives, i.e. the non-relevant images that have been classified as relevant.

Table 3
Accuracy of results.

Concepts	Visual Accuracy	Textual Accuracy	Hybrid Accuracy
Cleat	91.06%	94.44%	94.93%
Ski boot	94.69%	95.17%	97.10%
High_heel	93.96%	93.00%	96.86%
Lacing closure	92.27%	91.06%	93.72%
Heel with spring	93%	95.65%	94.44%
Tongue	92.75%	98.07%	97.10%
Toe caps	91.55%	94.20%	94.20%
Roller skates	90.10%	95.17%	96.86%
Average	92.42%	94.60%	95.65%

⁴ The dataset can be downloaded at: http://mklab.iti.gr/files/concepts-patent_images.rar.

Table 4Precision, recall and *F*-score for the concept detectors.

Concepts	Visual			Textual			Hybrid		
	Prec.	Recall	<i>F</i> -score	Prec.	Recall	<i>F</i> -score	Prec.	Recall	<i>F</i> -score
Cleat	84.38%	45.76%	59.34%	89.13%	69.49%	78.10%	89.58%	72.88%	80.37%
Ski boot	84.62%	67.35%	75.00%	87.18%	69.39%	77.27%	93.02%	81.63%	86.96%
High heel	82.69%	72.88%	77.48%	76.79%	72.88%	74.78%	92.59%	84.75%	88.50%
Lacing closure	79.17%	41.30%	54.29%	63.64%	45.65%	53.16%	88.46%	50.00%	63.89%
Heel with spring	69.70%	54.76%	61.33%	96.15%	59.52%	73.53%	100%	45.24%	62.30%
Tongue	75.68%	57.14%	65.12%	100%	83.67%	91.11%	95.12%	79.59%	86.67%
Toe caps	60.53%	53.49%	56.79%	75.68%	65.12%	70.00%	70.21%	76.74%	73.33%
Roller skates	82.50%	49.25%	61.68%	86.15%	83.58%	84.85%	96.55%	83.58%	89.60%
Average	77.41%	55.24%	63.88%	84.34%	68.66%	75.35%	90.69%	71.80%	78.95%

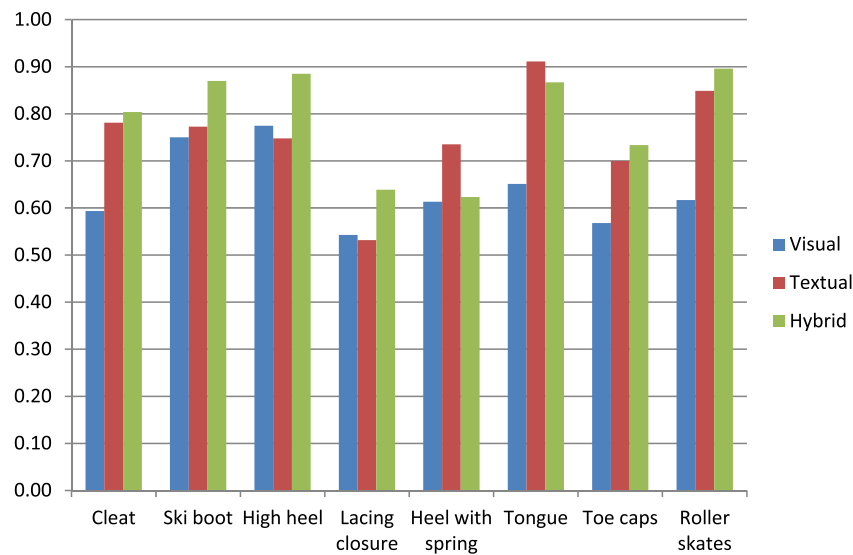
**Fig. 5.** *F*-score achieved by the employed concept-detectors.**Fig. 6.** Results for "ski-boots" using textual features. The green ticks indicate that the results retrieved are correct while the red × indicates the wrong results. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



Fig. 7. Results for “ski-boots” using visual features. The green tics indicate that the results retrieved are correct while the red × indicates the wrong results. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

FN = False Negatives, i.e. the relevant images that have been classified as non-relevant.

$$\text{Recall} = \frac{TP}{TP + FN}$$

In addition, we calculate *F*-score, precision and recall as follows:

$$F = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

6.2. Quantitative evaluation

In this section we present the results after realizing the training cases discussed in Section 5.2. Table 3 reports the accuracy for each concept detector and each training case.

Taking a first look on the results it seems that the accuracy of the three approaches is very high for all concepts. In most of the cases

in which, precision and recall are defined as follows:

$$\text{precision} = \frac{TP}{TP + FP}$$



Fig. 8. Results for “ski-boots” using the hybrid approach. The green tics indicate that the results retrieved are correct. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



Fig. 9. Results for query by visual example when the query is a “ski boot”. The green ticks indicate that the results retrieved are correct, while the red × indicates the wrong results. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the best performance is demonstrated by the hybrid approach, while the textual features seem to outperform the visual ones.

In order to have a better insight of the results from the information retrieval perspective we also report the precision, recall and *F*-score metrics (Table 4).

In this comparison we consider the *F*-score as the most important metric to represent the performance as it depends both on precision and recall. Of course depending on the application the classifiers could be optimised to maximise the recall or the precision. Considering the average *F*-scores, the visual-based training achieves a 63.88%, while when the textual features are employed the performance is increased to a 75.35%. Finally the hybrid approach demonstrates a slightly more improved *F*-score reaching 78.95%.

In Fig. 5 we present a graph view of the *F*-score performance for each concept detector. It is clear that in most cases (six of the eight) the hybrid approach performs better compared to the other two. However, the textual features outperform the hybrid training for the concepts “heel with spring” and “tongue”. On the other hand, textual-based results outperform the visual ones with only “high heels” and “lacing closure” being the exceptions. In the next section we will discuss in detail and provide examples that demonstrate for which cases the visual and the textual-based approaches provide quality results and for which they could fail.

6.3. Presentation of visual results

In order to provide some insights regarding the performance of the proposed approach, we will illustrate visual results of the concept detection through interaction modes of PatMedia.⁵ In this scenario we assume that the user is interested in retrieving figures depicting “ski boots”. We will present three cases, in which first the textual, second the visual and finally the hybrid approach are employed in the same dataset.

In Fig. 6 we present the first 18 results of the “ski boot” concept detector using only textual features. In this case the precision

achieved is 88.89% (16/18). It is interesting, however to see why the detector has failed in the last two figures. By examining the captions we notice: “Fig. 5 is a rear view of Fig. 2” and “Fig. 14 is a front view of the snowboard boot depicted in Fig. 13”. In the first case it is very reasonable that the current textual approach has failed since the description provides a reference to another figure. This could be handled with a more semantic analysis approach, in which we could replace the figure reference with the corresponding description to improve the results. On the other hand, the second caption can be considered misleading as it describes a snowboard boot, although the image describes the lacing system of a boot and therefore it was annotated as “lacing”.

In Fig. 7, we present the first 18 results, when only visual features are employed. In this case the reported precision is 94.44% (17/18). The first figure (top left), which is a lacing system, is not classified correctly due to the fact it resembles very much the part of a ski boot. However, when we combine both approaches (i.e. visual and textual), the results are improved as we achieve a 100% precision in the first 18 images (Fig. 8). This means that the 3 erroneous results appeared in the previous two approaches have been assigned a lower score.

Finally, it would be interesting to make a comparison with the query by visual similarity and discuss the results. It is important to state that query by visual similarity serves a different purpose, which is to retrieve images that look very similar, while in the case of concept detection the user looks for a specific concept, which could have several visual and textual representations with common characteristics. Taking a look at the results after submitting a query by having a ski boot as visual example, we indeed retrieve very similar images. However the system still retrieves another 3 images that are not ski-boots. This leads to a precision of 83.3% (15/18) as we can see in Fig. 9. Another important aspect to notice is that in this case only the very similar ski-boots were retrieved, while in the case of concept detection using the visual features, a variance of ski-boots with different orientations, perspectives etc. were retrieved due to the training with several ski boots examples.

Based on the aforementioned results it is obvious that the combination of visual and textual information performs better when compared to each single modality. Usually the text-based classification provides very good results, however the visual

⁵ A live demonstration of PatMedia is available at: <http://mklab.itl.gr/content/patmedia>.

classification is also very satisfactory and in several cases necessary in order to correct and complete unavoidable text processing limitations. Discussing the lessons learnt during these experiments it seems that on the one hand, the text-based classification could fail when the textual description is not clear or it is incomplete, while on the other hand, visual based classification could fail when two visually similar images are described with different concepts.

7. Conclusions

This article describes an attempt to build concept detectors for patent images combining visual and textual information using supervised machine learning and image analysis techniques. Although the experiments were conducted on a small dataset compared to the current databases of patent images, they can still provide feedback for some first conclusions. It seems that visual based classification can work complementarily to text classification results, but it can still have an acceptable performance in cases where the textual description is not available or incomplete. For instance, there are many patent documents where the textual descriptions cannot automatically be assigned to the correct figures or they cannot be automatically translated when they are written in certain foreign languages. On the other hand, all the image processing approaches (including this one) require prior segmentation of the drawing section pages to figures. Therefore, either automatic segmentation techniques could be applied, introducing, however, an error of around 20%, or manual segmentation, which is expensive in terms of time and human effort, could be performed. Another requirement of this method is that it needs a training set and a manual selection of concepts, while for each new concept introduced, there is a need to have manually annotated images by experts. The aforementioned constraints could be considered as significant obstacles for the scalability of the proposed method. To overcome these constraints, the following approaches can be considered. First, the drawings' page segmentation to figures could be performed by applying automatic approaches (e.g. [3]) and further improve the segmentation performance by considering supervised techniques (since segmentation in [3] is based on an unsupervised approach) as specifically trained OCR, in order to reduce the introduced error. Second, the concept selection process can be semi-automated by exploiting the IPC grouping information (e.g. as we see in Table 1, the selected concepts describe already specific groups and subgroups) and other external resources (i.e. taxonomies, ontologies). Finally, the construction of training sets for such frameworks can be performed either by collaborative annotation initiatives or even in an automatic way by exploiting user (i.e. patent searcher) implicit feedback (e.g. mouse clicks and keyboard inputs) during search. Recent approaches [9] have already provided promising results in the more generic area of image annotation.

From an application point of view it should be said that the concept retrieval module could be a part of a larger patent retrieval framework, which already includes functionalities such as full text and semantic search for the whole patent document. An additional important conclusion is the good performance of the AHDH feature in a supervised concept extraction framework, given the fact that the previous approaches and applications of AHDH have investigated only unsupervised image retrieval [10,32].

Future work includes providing results for more concepts and application of the same framework to larger datasets and to more IPC classes. Another challenge would be to test the performance of the detectors in the case of automatically segmented images (i.e. of lower quality). Finally, we plan to apply more intelligent combination and fusion techniques of textual and visual information, since the proposed approach adopted a rather simplistic technique (i.e. descriptor concatenation) for combining heterogeneous data.

This includes the investigation of late fusion and kernel fusion techniques, as well as the introduction of weights for the different modalities that could depend on the type of concept.

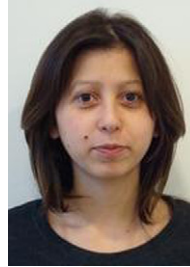
Acknowledgements

This work was supported by the project PESCaDO (FP7-248594) funded by the European Commission. Parts of this work were presented in IRFS 2011 [33]. The authors would like to thank Dominic De Marco for his help defining the concepts for patent images and Panagiotis Sidiropoulos for providing the code for AHDH feature extraction.

References

- [1] Zeng Z, Zhao J, Xu B. An outward-appearance patent-image retrieval approach based on the contour-description matrix. In: Proceedings of the 2007 Japan-China joint workshop on frontier of computer science and technology, p. 86–99; 2007.
- [2] Codina J, Pianta E, Vrochidis S, Papadopoulos S. Integration of semantic, metadata and image search engines with a text search engine for patent retrieval. Semantic search 2008 workshop, Tenerife, Spain, June 2, 2008.
- [3] Vrochidis S, Papadopoulos S, Moutzidou A, Sidiropoulos P, Pianta E, Kompatsiaris I. Towards content-based patent image retrieval; a framework perspective. World Patent Information Journal June 2010;32(2):94–106.
- [4] Tiwari A, Bansal V. PATSEEK: content based image retrieval system for patent database. In: Proceedings international conference on electronic business-04. Beijing, China: Tsinghua University; 2004.
- [5] Adams S. Electronic non-text material in patent applications – some questions for patent offices, applicants and searchers. World Patent Information 2005;27:99–103.
- [6] Ypma G. Evaluation of patent image retrieval. In: Information retrieval facility symposium 2010 (IRFS 2010), Vienna, Austria; June 1–4, 2010.
- [7] Yan R, Hsu W. Recent developments in content-based and concept-based image/video retrieval. In: Proceedings of the 16th ACM international conference on Multimedia (MM '08), New York, USA; 2008.
- [8] Giereth M, Koch S, Kompatsiaris Y, Papadopoulos S, Pianta E, Serafini L, Wanner L. A Modular framework for ontology-based representation of patent information. In: JURIX 2007: the 20th anniversary international conference on legal knowledge and information systems, Leyden, Netherlands; 2007.
- [9] Tsikrika T, Diou C, de Vries AP, Delopoulos A. Reliability and effectiveness of clickthrough data for automatic image annotation. Multimedia Tools and Applications 2011;55:27–52. Special Issue on Image and Video Retrieval: Theory and Applications, Springer.
- [10] Sidiropoulos P, Vrochidis S, Kompatsiaris I. Content-based binary image retrieval using the adaptive hierarchical dense histogram. Pattern Recognition Journal April 2011;44(4):739–50. Elsevier.
- [11] List J. How drawings could enhance retrieval in mechanical and device patent searching. World Patent Information 2007;29:210–8.
- [12] Sebe N, Lew MS, Zhou X, Huang TS, Bakker EM. The state of the art in image and video retrieval. In: Proceedings of the 2nd international conference on image and video retrieval, Urbana; July 2003.
- [13] Snoek CGM, Worring M. Multimodal video indexing, a review of the state-of-the-art. Multimedia Tools and Applications January 2005:5–35. Springer Netherlands.
- [14] Gkalelis N, Mezaris V, Kompatsiaris I. High-level event detection in video exploiting discriminant concepts. In: Proceedings of the 9th international workshop on content-based multimedia indexing (CBMI 2011), Madrid, Spain; June 2011.
- [15] Huiskes MJ, Thomee B, Lew MS. New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative in MIR '10. In: Proceedings of the international conference on multimedia information retrieval. New York, NY, USA: ACM; 2010.
- [16] Wei X-Y, Jiang Y-G, Ngo C-W. Concept-driven multi-modality fusion for video search. IEEE Transactions on Circuits and Systems for Video Technology 2011; 21(10):62–73.
- [17] Eakins JP. Trademark image retrieval. In: Lew M, editor. Principles of visual information retrieval. Berlin: Springer-Verlag; 2001.
- [18] Wu JK, Lam CP, Mehre BM, Gao YJ, Desai Narasimhalu A. Content-based retrieval for trademark registration. Multimedia Tools and Applications 1996;3:245–67.
- [19] LTU Technologies. <http://www.luttech.com/en/>.
- [20] eMARKS project. <http://emarks.iisa-innov.com/>.
- [21] Vrochidis S. Patent image retrieval. In: Information retrieval facility symposium 2008 (IRFS 2008), Vienna, Austria; November 5–7, 2008.
- [22] PATExpert (FP6-028116). <http://www.patexpert.org/>.
- [23] Mörzinger R, Horti A, Thallinger G, Bhatti N, Hanbury A. Classifying patent images. In: Proceedings of CLEF 2011, Amsterdam; September 19–22, 2011.
- [24] Csurka G, Renders J, Jacquet G. XRCE's participation at patent image classification and image-based patent retrieval tasks of the Clef-IP 2011. In: Proceedings of CLEF 2011, Amsterdam; September 19–22, 2011.
- [25] De Marco D. Mechanical patent searching: a moving target. Baltimore, USA: Patent Information Users Group (PIUG); May 1–6, 2010.

- [26] Hastie T, Tibshirani R. Discriminant adaptive nearest neighbor classification. *IEEE Pattern Analysis and Machine Intelligence (PAMI)* 1996;18:607–16.
- [27] Fischler MA, Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* 1981;24(6):381–95.
- [28] Porter MF. An algorithm for suffix stripping. *Program* 1980;14(3):130–7.
- [29] The lemur toolkit. <http://www.cs.cmu.edu/~lemur/> [lemur].
- [30] Boser BE, Guyon IM, Va VN. A training algorithm for optimal margin classifiers. In: COLT '92: proceedings of the fifth annual workshop on computational learning theory. New York, NY, USA: ACM Press; 1992. p. 144–52.
- [31] Chang C, Lin C. LIBSVM: a library for support vector machines. Software Available at: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [32] Sidiropoulos P, Vrochidis S, Kompatsiaris I. Content-based retrieval of complex binary images. In: Proceedings of the 8th international workshop on content-based multimedia indexing (CBMI 2010), p. 182–7, Grenoble, France; June 23–25, 2010.
- [33] Vrochidis S. Patent image retrieval based on concept extraction and classification. IRFS; 2011.



Anastasia Mourtzidou received the Diploma degree in electrical and computer engineering and two MSc degrees in Advanced Computer and Communication Systems and Informatics and Management, all from the Aristotle University of Thessaloniki in 2006, 2009 and 2011, respectively. Currently, she is a Research Associate with the Informatics and Telematics Institute – Centre for Research and Technology Hellas. Her research interests include semantic multimedia analysis, search engine development and patent retrieval. She has participated in many European projects related to semantic analysis and retrieval, and patent search. She is the coauthor of 1 article in refereed journal and 9 papers in international conferences.



Stefanos Vrochidis received the Diploma degree in Electrical Engineering from Aristotle University of Thessaloniki, Greece and the MSc degree in Radio Frequency Communication Systems from University of Southampton, in 2000 and 2001, respectively. Currently, he is an Associated Researcher at the Informatics and Telematics Institute. His research interests include semantic multimedia analysis retrieval, search engines, patent image search and human interactions, as well as environmental applications. He has successfully participated in many European and National projects and he has been involved as a coauthor of 6 articles in refereed journals, 2 book chapters and more than 20 papers in international conferences.



Ioannis (Yiannis) Kompatsiaris is a Senior Researcher with the Informatics and Telematics Institute leading the Multimedia Group. His research interests include semantic multimedia analysis, indexing and retrieval, Web 2.0 content analysis, knowledge structures, reasoning and personalization for multimedia applications. He received his Ph.D. degree in 3-D model based image sequence coding from the Aristotle University of Thessaloniki in 2001. He is the coauthor of 40 papers in refereed journals, 20 book chapters, 4 patents and more than 150 papers in international conferences. He is a Senior Member of IEEE and a member of ACM.