A Multimodal Approach for Semantic Patent Image Retrieval

Kader Pustu-Iren
TIB – Leibniz Information Centre for Science and Technology
Hannover, Germany
kader.pustu@tib.eu

Gerrit Bruns
TIB – Leibniz Information Centre for Science and Technology
Hannover, Germany
gerrit.bruns@tib.eu

Ralph Ewerth*
TIB – Leibniz Information Centre for Science and Technology
Hannover, Germany
ralph.ewerth@tib.eu

ABSTRACT

Patent images such as technical drawings contain valuable information and are frequently used by experts to compare patents. However, current approaches to patent information retrieval are largely focused on textual information. Consequently, we review previous work on patent retrieval with a focus on illustrations in figures. In this paper, we report on work in progress for a novel approach for patent image retrieval that uses deep multimodal features. Scene text spotting and optical character recognition are employed to extract numerals from an image to subsequently identify references to corresponding sentences in the patent document. Furthermore, we use a neural state-of-the-art CLIP model to extract structural features from illustrations and additionally derive textual features from the related patent text using a sentence transformer model. To fuse our multimodal features for similarity search we apply re-ranking according to averaged or maximum scores. In our experiments, we compare the impact of different modalities on the task of similarity search for patent images. The experimental results suggest that patent image retrieval can be successfully performed using the proposed feature sets, while the best results are achieved when combining the features of both modalities.

CCS CONCEPTS

• **Information systems** → **Image search**; **Content analysis and feature selection**; • **Computing methodologies** → **Visual content-based indexing and retrieval**; **Image representations**.

KEYWORDS

Patent Image Similarity Search, Deep Learning, Mulitmodal Feature Representations, Scene Text Spotting

1 INTRODUCTION

Patent experts and researchers often encounter language and terminology barriers when conducting searches to identify research or patent gaps, (newly) emerging technology developments, or to check the patentability of research results. Existing patent retrieval methods are primarily based on textual searches and largely exclude illustrations and the relationship between text and image. Often, however, the innovation of a patent can be identified with the help of an illustration, and patents with similar or related innovations can be quickly analysed by looking at illustrations in a comparative way. In this context, a survey with patent experts confirms the importance of illustrations in their high informative value and the demand for an image-based search [8]. Moreover, with the continuous refinement of already patented research, the terminology used changes [3], making it more difficult to find corresponding patents. This problem is exacerbated when cross-linguistic searches are conducted. Therefore, illustrations provide an alternative way to enable the identification of relevant results in patents, regardless of language and terminology. The use of illustrations is also advantageous for domain and patent class independent searches. In this way, intellectual property (IP) rights can be evaluated for further application domains, which is only possible to a limited extent with a purely textual search. This is especially relevant for basic and technical patents, whose scope of application is often not clear at the beginning of the creation of an exploitation strategy.

In this paper, we present a novel multimodal system for semantic patent image retrieval in a query-by-example scenario. To extract visually relevant features from images, pre-trained embeddings using deep neural networks are used. Furthermore, scene text spotting is applied in order to extract numerals from the images and map them to their mentions in the patent text. Next, we derive textual features from the relevant sentences in the text utilizing sentence transformers. Finally, textual and visual features are used to index the represented illustrations. Experimental results are presented for semantic image search investigating both unimodal and multimodal feature sets.

The rest of the paper is organized as follows. We review related work in Section 2. Section 3 introduces the proposed approach for multimodal patent image search. We provide an experimental evaluation of the proposed solution in Section 4 and conclude the paper with a short discussion of results in Section 5.

2 RELATED WORK

Previous approaches to patent information retrieval have been largely limited to textual information [19]. However, terminology in patents changes continuously due to the constant evolution of the presented content and is inconsistent for this reason [3]. Often, innovative terminology is "invented" along with the actual invention. One result of this evolution is that search results are often incomplete and do not display all relevant patents. The (additional) evaluation of non-textual information in the form of illustrations, such as technical drawings, graphs and diagrams, can facilitate and significantly improve the search for similar or relevant patents. In addition, references to the relevant text passages are often given in numerical form in these illustrations, so that automatic recognition of these image-text references can also significantly improve the quality of the (multimodal) search results.

*Also with L3S Research Center, Leibniz University Hannover, Germany.

PatentSemTech, July 15th, 2021, online

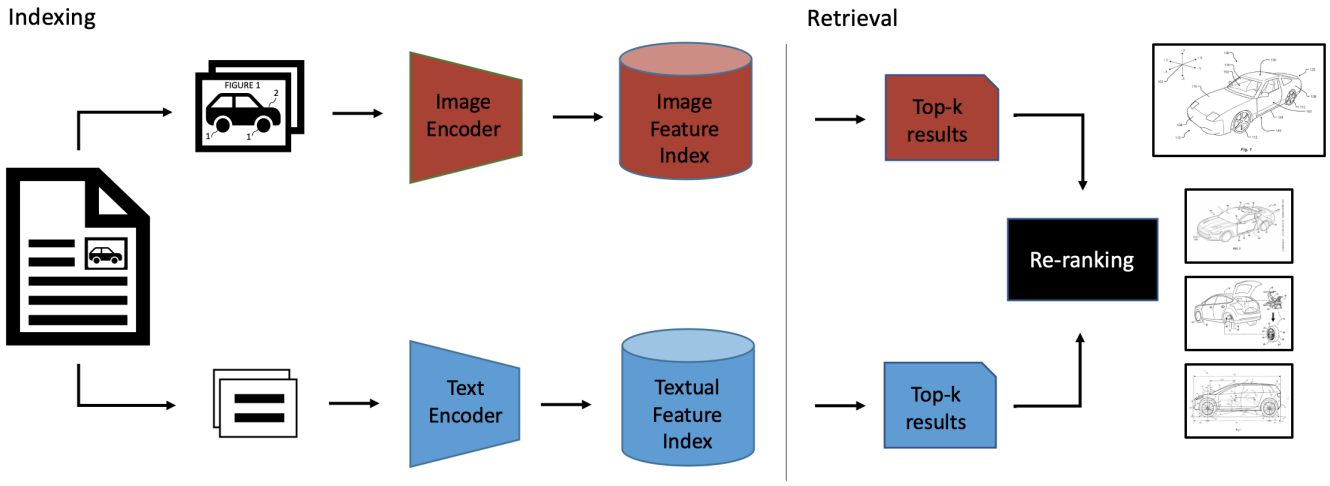


Figure 1: Proposed system for multimodal patent image retrieval.

The more general problem of searching in image databases (image retrieval) has been intensively researched in the last decades. Simpler methods for search in image databases are usually based on so-called low-level features, which technical descriptions of shape, color, or texture. However, results based on such features very often do not meet the search needs of users, which are mostly of a content or semantic nature ("semantic gap") [21]. In recent years, significant progress has been made to automatically recognize content in images (denoted as object recognition or visual concept detection) [22], especially through deep learning approaches [5, 10, 30]. In this way, search queries of a content-related nature can be more accurately answered.

An important aspect of the presented approach is the similarity search that follows feature extraction. Current similarity search approaches learn compact codes to replace images [18, 27, 28]. The compact codes usually compress high-dimensional features of a Convolutional Neural Network (CNN) trained on specific datasets suitable for the given task. However, these methods are not optimized for the technical and schematic illustrations in patents, so there is a need for research and development in this area.

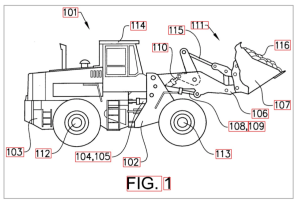
So far, there are relatively few specific approaches for searching visual information in patents [29]. An example is the Patmedia method for similarity search [25], extensions of this [20, 23, 24], or other approaches for concept-based graphical search [11, 13]. These methods generally extract textual and visual low-level features from patent images and train detectors that identify a limited number of predefined concepts. Experiments of these works show that the combination of visual and textual features works best for the task of concept detection. More recent approaches [9, 14] establish the references of figures and related text passages using an automatic detection of the corresponding numerical referencing in the figures. Another approach [4] uses SIFT-like local histograms as features and represents the images in patents using Fisher vectors. In the experiments based on the 2011 CLEF-IP evaluation [15], the best retrieval results were achieved by late fusion of textual and non-textual results. Bhatti and Hanbury [3] provide an overview of further research regarding specific figure types (photo, flow chart, technical drawings, diagrams, graphs) that may also be relevant for patent retrieval. However, to date no integrated patent retrieval system exploiting multimodal search does exist. The representation and quality of the images in patents as well as their schematic and sketchy character require specific approaches or the recognition of special objects that are particularly relevant in patents.

3 MULTIMODAL PATENT IMAGE SEARCH

We now discuss the proposed system that incorporates multimodal patent features to establish a similarity search based on illustrations. Figure 1 illustrates the individual steps. First, we extract visual and textual features (Section 3.1, 3.2) from the patent images. Then, based on each modality an index of corresponding image feature vectors is built (Section 3.3). Finally, the most similar results to a query image can be retrieved by re-ranking results based on both indexes.

3.1 Image Feature Extraction

Patent images are a special category of images that have sketch-like characteristics. They usually consist of technical drawings, diagrams, or graphs and are mostly black and white. While smaller details can often be of great relevance for interpretation, they often also contain redundant patterns. To represent these kind of images, features are extracted using deep neural network. We use the Contrastive Language-Image Pre-training (CLIP) [16] model that was trained on a multimodal dataset of 400 million image-text-pairs collected from the internet. The CLIP model is aimed at learning visual concepts from natural language supervision and is primarily designed for flexible zero-shot computer vision classification on arbitrary image datasets by providing simple textual image descriptions. This powerful approach has improved the state of the art on several benchmark datasets including ImageNet Sketch [26],



The wheel loader 101 comprises an equipment 111 for handling objects or material. The equipment 111 comprises a load-arm unit 106 and an implement 107 in the form of a bucket fitted on the load-arm unit.A first end of the load-arm unit 106 is pivotally connected to the front vehicle section 102. The implement 107 is pivotally connected to a second end of the load-arm unit 106. ...

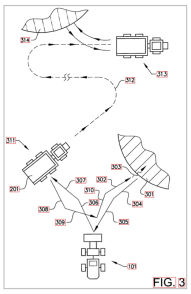


Figure 3 shows a driving pattern comprising a series of steps from excavation to loading onto the articulated hauler 201. Specifically, the wheel loader 101 travels forward, see arrow 302, to the natural ground 301 in a straight position, wherein the front and rear vehicle parts are in line. When it approaches the natural ground 301, it thrusts into the natural ground, see arrow 303. The lifting arm unit is raised, wherein the bucket 107 is filled with material from the natural ground. ...

Figure 2: Image-text relations through OCR.

which contains sketch images with characteristics similar to patent images. This motivates us to utilize CLIP embeddings for the task of patent image similarity search. In particular, we use the pre-trained vision transformer (ViT-B/32) to extract visual features and embed the patent images.

3.2 Textual Feature Extraction

Patent figures usually contain image text, particularly numbers that can be used to link illustrated concepts to a description in the patent document. To use these textual descriptions, we first apply scene text spotting methods (Section 3.2.1). After relevant sentences have been identified, they are embedded using sentence transformers (Section 3.2.2).

3.2.1 Image-Text Relations using OCR. Optical Character Recognition (OCR) aims to recognize characters in images. Recently, scene text recognition methods based on neural networks have emerged. We use a two-step approach in which we first detect text blocks and then recognize the text they contain. For scene text detection, the CRAFT (Character Region Awareness For Text detection) [2] model for character-level text detection is applied. Subsequently, a four-stage deep scene text recognition (STR) framework [1] is employed to extract the text. Although these methods were trained for recognizing text in real-world scenes, they prove to be very accurate on patent images, for which text detection and recognition is generally easier than for scene text. Once the image text is extracted, we keep the numbers and prune irrelevant text. The numbers are then used to identify the relevant sentences in the XML file of the corresponding patent document. For this purpose, we tokenize the text, search for the numbers and keep all matching sentences that provide a description for the illustrated concepts. Exemplary text mappings resulting from the scene text recognition can be seen in Figure 2.

3.2.2 Sentence Transformers. Sentence transformer neural networks were recently introduced and can be used to compute dense vector representations for sentences. We use a RoBERTa [12] model that was pre-trained to produce semantically meaningful sentence embeddings (accordingly to Sentence-BERT[17]) and optimized for semantic textual similarity (STS) in the English language. We embed all the sentences found in the previous image-text mapping step. Finally, an average vector over all related sentences is created to represent an patent image.

3.3 Similarity Search

Based on the extracted feature representations, indexes are built using the FAISS library[7]. An index is based on product quantization [6] and allows for the efficient comparisons between query vectors and stored vectors based on cosine similarity and returns nearest neighbors. We built separate indexes for both the image and textual feature modalities based on a dataset comprised of 30,379 patent images. Subsequently, the nearest neighbors of a query image can be retrieved by similarity search based a) on the stored visual features, b) on the stored textual features, or c) on the basis of a combination of ranking results of both indexes. For the last option we explore two different re-ranking approaches. The first one is based on averaging the resulting similarity scores of each modality, whereas in the second strategy the final ranking is based on reordering according to maximum scores.

4 EVALUATION AND DISCUSSION

In this section, the patent image retrieval approaches are evaluated according to the experimental setup in Section 4.2) and based on a predefined patent collection (Section 4.1). We discuss outcomes of the experiments in Section 4.3.

4.1 Patent Dataset

We conduct our retrieval experiments on a patent collection from the European Patent Office (EPO) focusing on the exemplary fields of autonomous driving and wind power. To this end, we download patents from the time period 2007 to 2020 and ensure that each patent contains an XML file to parse the structured text and image information. After excluding formulas, our final patent collection comprises 2,858 patent documents with a total of 30,379 figures of technical drawings, diagrams and graphs. Analogously, another 3,770 images from 300 patent documents are reserved as test data.

4.2 Experiments

The performance of our system is evaluated using the average precision (AP) score which is the most commonly used quality measure for retrieval approaches. The AP score is calculated from a list of ranked documents as follows:

$$\mathrm{AP} = \sum_n (R_n - R_{n-1}) P_n \tag{1}$$

where R_n and P_n are the precision and recall at the n th threshold. In general, AP is the average of the precision scores at each relevant document. To evaluate the overall performance, the mean AP (mAP) score is calculated by taking the mean value of the AP scores across different queries. To verify the performance of our system, we randomly selected 20 query images along with their descriptions (described in Section 3.2.1) from the test data and evaluated

Table 1: mAP results up to rank 50 for randomly chosen queries. Re-ranking (avg) denotes the averaging of the different modalities' scores. Re-ranking (max) denotes the reordering according to maximum scores.

Textual Features	Visual Features	Re-ranking	
		max	avg
0.683	0.696	0.703	0.715

AP scores for the textual retrieval, visual retrieval and combined retrieval based on re-ranking. To evaluate the relevance of an retrieval results we rely on the annotator assessment (done by one of the authors). Using the additional figure descriptions assists in evaluating the relevance of retrieval results to the query image. The ranked retrieval lists are evaluated for the top-50 ranks using the AP score (AP@50).

4.3 Discussion

The results of our experiments are shown in Table 1. Using only visual features for image retrieval yields a slightly higher mAP score of 0.696 compared to using textual features. The combination of both modalities yields the highest mAP score of 0.715 when scores of the textual and visual similarity search are averaged. Reordering the similarity values according to the maximum scores for both feature sets had a smaller effect on the similarity search performance. The results suggest that combining both modalities can help increase the quality of retrieval results. In general, results based on visual features were easier to annotate since the visual embeddings retrieve mostly visually similar results. It should also be noted that results based on textual features were harder to inspect and thus annotated with the additional help of the sentences representing the retrieved image. In general, it was observed that textual features retrieved semantically relevant images. Thus, the combination of both feature representations presents a good mixture of both visually and semantically related patent images.

5 CONCLUSIONS

The discussion of related work for patent image retrieval revealed that existing work is either outdated or insufficient when it comes to exploiting the multimodal information that patents provide. In this paper, we have presented a framework that exploits multimodal features to enable semantic patent image search. Image-text relations are identified through scene text spotting and OCR yielding a mapping of in-figure numbers to the corresponding text. This allowed us to embed relevant text passages in feature vector representations. Additionally, we successfully embedded the shape and topological information in images using powerful deep neural networks. We exploit both textual and image features in order to facilitate semantic similarity for patent images. Experimental results demonstrated the feasibility of the approach, while suggesting that the combination of both modalities is beneficial.

In the future, we plan to exploit further information in images such as non-numeric image text. Moreover, we plan to incorporate multimodal information in an end-to-end network and have a joint framework to conduct patent search. Thereby, we intent to fuse features by exploiting multimodal machine learning architectures.

ACKNOWLEDGEMENTS

We would like to sincerely thank the reviewers for their valuable and comprehensive comments. This work is financially supported by the Federal Ministry of Education and Research (BMBF, Bundesministerium für Bildung und Forschung, project reference 01IO2004A).

REFERENCES

[1] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. 2019. What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 4714–4722. https://doi.org/10.1109/ICCV.2019.00481

[2] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character Region Awareness for Text Detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 9365–9374. https://doi.org/10.1109/CVPR.2019.00959

[3] Naeem Bhatti and Allan Hanbury. 2013. Image search in patents: a review. *International Journal on Document Analysis and Recognition* 16, 4 (2013), 309–329. https://doi.org/10.1007/s10032-012-0197-5

[4] Gabriela Csurka, Jean-Michel Renders, and Guillaume Jacquet. 2011. XRCE's Participation at Patent Image Classification and Image-based Patent Retrieval Tasks of the Clef-IP 2011. In *CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands (CEUR Workshop Proceedings, Vol. 1177)*, Vivien Petras, Pamela Forner, and Paul D. Clough (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-1177/CLEF2011wn-CLEF-IP-CsurkaEt2011.pdf

[5] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[6] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. 2011. Product Quantization for Nearest Neighbor Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 1 (2011), 117–128. https://doi.org/10.1109/TPAMI.2010.57

[7] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *CoRR* abs/1702.08734 (2017). arXiv:1702.08734 http://arxiv.org/abs/1702.08734

[8] Hideo Joho, Leif Azzopardi, and Wim Vanderbauwhede. 2010. A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. In *Information Interaction in Context Symposium, IIiX 2010, New Brunswick, NJ, USA, August 18-21, 2010*, Nicholas J. Belkin and Diane Kelly (Eds.). ACM, 13–24. https://doi.org/10.1145/1840784.1840789

[9] R. Kramer and U. Döring. 2016. CLEF-IP 2011: Tool zur Unterstützung der bildorientierten Selektion von Patentdokumenten am Beispiel des XPAT Patent Viewers. In *Big Data - Chancen und Herausforderungen. 38. Kolloquium der Technischen Universität Ilmenau über Patentinformation und gewerblichen Rechtsschutz. Proceedings. PATINFO*. 209–2019.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (Eds.). 1106–1114. https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[11] Dimitris Liparas, Anastasia Moumtzidou, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2014. Concept-oriented labelling of patent images based on Random Forests and proximity-driven generation of synthetic data. In *Proceedings of the Third Workshop on Vision and Language, VL@COLING 2014, Dublin, Ireland, August 23, 2014*, Anja Belz, Darren Cosker, Frank Keller, William Smith, Kalina Bontcheva, Sien Moens, and Alan F. Smeaton (Eds.). Dublin City University and the Association for Computational Linguistics, 25–32. https://doi.org/10.3115/v1/W14-5404

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[13] Hui Ni, Zhenhua Guo, and Biqing Huang. 2015. Binary Patent Image Retrieval Using the Hierarchical Oriented Gradient Histogram. In *International Conference on Service Science, ICSS 2015, Weihai, Shandong, China, May 8-9, 2015*. IEEE Computer Society, 23–27. https://doi.org/10.1109/ICSS.2015.12

[14] Jeong Beom Park, Thomas Mandl, and Do Wan Kim. 2017. Patent Document Similarity Based on Image Analysis Using the SIFT-Algorithm and OCR-Text. *International Journal of Contents* 13(4) (2017), 70–79.

[15] Florina Piroi, Mihai Lupu, Allan Hanbury, and Veronika Zenz. 2011. CLEF-IP 2011: Retrieval in the Intellectual Property Domain. In *CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands (CEUR Workshop Proceedings, Vol. 1177)*, Vivien Petras, Pamela Forner, and Paul D. Clough (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-1177/CLEF2011wn-CLEF-IP-PiroiEt2011.pdf

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *CoRR* abs/2103.00020 (2021). arXiv:2103.00020 https://arxiv.org/abs/2103.00020

[17] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. https://doi.org/10.18653/v1/D19-1410

[18] Josiane Rodrigues, Marco Cristo, and Juan G Colonna. 2020. Deep hashing for multi-label image retrieval: a survey. *Artificial Intelligence Review* 53, 7 (2020), 5261–5307.

[19] Walid Shalaby and Wlodek Zadrozny. 2019. Patent retrieval: a literature review. *Knowl. Inf. Syst.* 61, 2 (2019), 631–660. https://doi.org/10.1007/s10115-018-1322-7

[20] Panagiotis Sidiropoulos, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2011. Content-based binary image retrieval using the adaptive hierarchical density histogram. *Pattern Recognition* 44, 4 (2011), 739–750. https://doi.org/10.1016/j.patcog.2010.09.014

[21] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 12 (2000), 1349–1380.

[22] Cees G. M. Snoek and Arnold W. M. Smeulders. 2010. Visual-Concept Search Solved? *Computer* 43, 6 (2010), 76–78. https://doi.org/10.1109/MC.2010.183

[23] Stefanos Vrochidis, Anastasia Moumtzidou, and Ioannis Kompatsiaris. 2012. Concept-based patent image retrieval. *World Patent Information* 34 (2012), 292–303.

[24] Stefanos Vrochidis, Anastasia Moumtzidou, and Ioannis Kompatsiaris. 2014. Enhancing Patent Search with Content-Based Image Retrieval. In *Professional Search in the Modern World - COST Action IC1002 on Multilingual and Multifaceted Interactive Information Access*, Georgios Paltoglou, Fernando Loizides, and Preben Hansen (Eds.). Lecture Notes in Computer Science, Vol. 8830. Springer, 250–273. https://doi.org/10.1007/978-3-319-12511-4_12

[25] Stefanos Vrochidis, S. Papadopoulos, Anastasia Moumtzidou, Panagiotis Sidiropoulos, Emanuelle Pianta, and Ioannis Kompatsiaris. 2010. Towards content-based patent image retrieval: A framework perspective. *World Patent Information* 32 (2010), 94–106.

[26] Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. 2019. Learning Robust Global Representations by Penalizing Local Predictive Power. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 10506–10518. https://proceedings.neurips.cc/paper/2019/hash/3eefceb8087e964f89c2d59e8a249915-Abstract.html

[27] Jun Wang, Wei Liu, Sanjiv Kumar, and Shih-Fu Chang. 2015. Learning to hash for indexing big data - A survey. *Proc. IEEE* 104, 1 (2015), 34–57.

[28] Jingdong Wang, Ting Zhang, Nicu Sebe, Heng Tao Shen, et al. 2017. A survey on learning to hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 4 (2017), 769–790.

[29] Liping Yang, Ming Gong, and Vijayan K. Asari. 2020. Diagram Image Retrieval and Analysis: Challenges and Opportunities. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*. IEEE, 685–698. https://doi.org/10.1109/CVPRW50498.2020.00098

[30] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. 2018. Learning Transferable Architectures for Scalable Image Recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 8697–8710. https://doi.org/10.1109/CVPR.2018.00907