Visual-based classification of figures from scientific literature

Theodoros Giannakopoulos, Ioannis Foufoulas, Eleftherios Stamatogiannakis, Harry Dimitropoulos, Natalia Manola, and Yannis Ioannidis
Management of Data, Information, and Knowledge Group, University of Athens, Greece
{tyiannak, johnfouf, estama, harryd, natalia, yannis}@di.uoa.gr

ABSTRACT

Authors of scientific publications and books use images to present a wide spectrum of information. Despite the richness of the visual content of scientific publications the figures are usually not taken into consideration in the context of text mining methodologies towards the automatic indexing and retrieval of scientific corpora. In this work, we present a system for automatic categorization of figures from scientific literature to a set of predefined classes. We have employed a wide range of visual features that achieve high discrimination ability between the adopted classes. A real-world dataset has been compiled and annotated in order to train and evaluate the proposed method using three different classification schemata.

Categories and Subject Descriptors

I.4.9 [**IMAGE PROCESSING AND COMPUTER VISION**]: Applications

1. INTRODUCTION

During the last years document analysis and text mining techniques have been widely applied in scientific documents. In addition, while Open Access is becoming rapidly widespread and through dynamic and online document formats, more visual information is also available. Despite that huge increase in visual content, the application of computer vision techniques in images from scientific publications have not yet reached a high maturity level or an admissible level of automation. During the last years, a few scientific image search engines have been made available, however they mainly focus on extracting information from the respective image captions through typical text mining procedures. Most of the related existing research efforts focus on biomedical images in the context of the spreading biomedical literature mining [6, 10]. In [5] a method towards automatic figure categorization in the context of a digital library is presented, while [3] proposes a rule-based methodology for segmenting a document into textual and non-textual blocks. In this work, we present a method for scientific figure classification adopting a wider range of visual features, focusing to achieve increased discrimination ability in terms of the particular classification task under study. The figures are automatically classified to five general categories which cover the largest part of the interdisciplinary scientific fields. A real-world dataset and extensive experimental results are also presented in order to prove the validity of the method. The proposed system can be either used as an automatic labeling approach for scientific figures or as a method for image similarity calculation in the context of a retrieval system.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742024.

2. DATASET

A real-world dataset has been compiled in order to train and evaluate the proposed scheme. Almost 1500 images have been manually labeled to the following categories: (a) Charts (2D and 3D plots) (b) Diagrams (c) Geometric shapes and visualizations (2D and 3D) (d) Maps and continuous 2D representations (e.g. medical images, signal spectrograms) and (e) Photographs. The images have been extracted from more than 500 scientific publications from the `arXiv` dataset (`http://arxiv.org/`), covering a wide range of content classes from several scientific domains. This work only focuses on image classification - not segmentation - therefore the images of the particular dataset are of homogeneous content.

3. VISUAL FEATURE EXTRACTION

The adopted visual features used to represent each image in the context of an effort to produce a class-discriminant feature space are the following:

- **Color:** We extract color-related histogram features from each image, in particular 8-bin histograms of the following channels: red, blue, green, grayscale and saturation.
- **Edges:** A normalized histogram of the edge Sobel operator is computed over the grayscale values.
- **Lines:** Lines can be very informative in discriminating between diagrams, charts and other types of figures. We use the Canny detector to detect edges along with the Hough transform for detecting lines [2]. Three line-related statistics are finally extracted for each image.
- **Histograms of Oriented Gradients:** HOGs represent an object using the local distributions of intensity

Table 1: Overall Performance results (F1)

kNN	SVMs	DBN
70.9%	75.4%	76.7%

gradients and edge directions. They have been widely used in object and human tracking [1].

- **Local binary patterns:** LBPs [7] form a widely used feature in modern image analysis methods. In general, LBPs endode local pixel neighborhoods using binary representations, hence their name. We have selected to adopt LBPs for their ability to represent differences in texture characteristics between images.

- **Face-related attributes:** In order to qualify the existence of faces, the Viola-Jones face detector has been applied on each image, and the number of faces along with three bounding box-related statistcs are finally used as features.

- **Text-related attributes:** The Tesseract Optical Character Recognition (OCR) engine [8] is used in order to extract textual information from the figures. Then the following statistics are extracted as features: (a) number of words, (b) characters per word ratio and (c) percentage of numbers.

4. CLASSIFICATION

We have implemented and evaluated the following types of classification methodologies: (a) The k-nearest neighbor classifier (kNN), which, despite its simplicity, is widely used both for binary and multi-class tasks. It does not actually require a training stage and it can operate directly in a multi-class mode (as the one required in the current work) [9]. (b) Support Vector Machines (SVMs) are state-of-the-art classifiers widely employed in many machine learning applications [9], making use of supporting hyperplanes parallel to the decision. (c) Deep Belief Networks have also gained much research attention in the field of computer vision. They are based on the idea that Restricted Boltzmann Machines (RBMs) can be stacked to form a hierarchy of multiple layers where the output of each RBM is used as input to the next [4].

5. EXPERIMENTAL EVALUATION

Table 1 presents the overall performances (F1 measure) for all three classification methods, proving that the Deep Belief Network method outperforms the rest. In addition, Table 2 shows the confusion matrix of this method along with the recall, precision and F1 performance measures. In order to achieve statistical accuracy of the estimated performance measures, repeated cross-validation has been followed during the experimentation.

6. CONCLUSIONS

We presented a wide range of visual features that discriminate between different types of figures from scientific literature and three classification approaches which have been evaluated on a real-world dataset. The achieved classification performance is promising if we take into consideration the fact that this work focuses on generalized class definitions. Our ongoing and future work focuses on defining a hierarchical classification that provides deeper information regarding the types of images and evaluating the adopted feature space in the context of a scientific image retrieval system, in terms of how accurately can it correlate to semantic distances between scientific visual content.

Table 2: Detailed performance measures for DBN

	CH	DI	GE	MA	PH	Rec	Pre	F1
CH	79	12	5	3	1	79	80	80
DI	10	74	11	3	1	75	71	73
GE	5	13	68	9	3	68	71	70
MA	3	3	8	74	11	74	76	76
PH	0	1	3	8	87	87	85	86

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the EU's Horizon2020 (OpenAIRE2020 project).

8. REFERENCES

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[2] R. O. Duda and P. E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972.

[3] J. L. Fisher, S. C. Hinds, and D. P. D'Amato. A rule-based system for document image segmentation. In *Pattern Recognition, 1990. Proceedings., 10th International Conference on*, volume 1, pages 567–572. IEEE, 1990.

[4] G. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[5] X. Lu, S. Kataria, W. J. Brouwer, J. Z. Wang, P. Mitra, and C. L. Giles. Automated analysis of images in documents for intelligent document search. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(2):65–81, 2009.

[6] H. Müller, J. Kalpathy-Cramer, D. Demner-Fushman, and S. Antani. Creating a classification of image types in the medical literature for visual categorization. In *SPIE medical imaging*, pages 83190P–83190P. International Society for Optics and Photonics, 2012.

[7] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[8] R. Smith. An overview of the tesseract ocr engine. In *ICDAR*, volume 7, pages 629–633, 2007.

[9] S. Theodoridis and K. Koutroumbas. *Pattern Recognition, Fourth Edition*. Academic Press, 2008.

[10] A. D. Yuan X. A novel figure panel classification and extraction method for document image understanding. *International Journal of Data Mining and Bioinformatics*, 9(1):22–36, 2014.