Extending Document Exploration with Image Retrieval: Concept and First Results

Lin Shao, Mathias Glatz, Eric Gergely, Markus Müller, Denis Munter, Stefan Papst and Tobias Schreck

Graz University of Technology, Austria

Abstract

Information retrieval provides to date effective methods to search for documents relevant to user queries, and to support exploration of clusters of similar documents. Typically, the retrieval relies on text-based queries and similarity functions. However, in many cases also visual content is important in documents, for example, in the visualization field. There, researchers may want to search for papers based on similar example visualizations, which is difficult by relying on keyword search alone. We present a concept to automatically label visualization types in research papers and search for similar images, relying on state of the art image descriptors. We created a prototype that allows to search for papers showing images similar to a query image. Preliminary results of applying it on a corpus of VAST papers indicate the chosen descriptors can retrieve papers with similar images. Our approach for image-based search can complement text-based search and in perspective, support document corpus exploration based on clustering contained images. In future work, we want to explore if image-based search can also support the formation of taxonomies of a corpus or research papers, based on image similarity.

CCS Concepts

•***Information systems*** → *Image search;* •***Computing methodologies*** → *Visual content-based indexing and retrieval;*

1. Introduction

Students and researchers are often interested in related works that specifically address their particular areas and research topics. For instance, researchers in the visualization community may be interested in visual results/contributions of a paper. Often slightly related approaches, e.g., part of a technique, application area or alternative representation, may already be interesting. The problem of the common basic textual search is that such smaller contributions, which are not indexed or strongly stressed in the paper, can be hard to find. These search techniques typically use textual annotations such as keywords, title, abstract or other meta-data. In addition, the number of keywords are often limited from a catalog. Existing document visualization techniques help users to identify and compare topics of interest in documents, e.g., [CGS*11, LZP*12]. In [SOR*09], a visual abstraction of documents by a word cloud approach integrating images is proposed. Typical document retrieval systems are text-based, and do not include image similarity measures, although many techniques exist for content-based image retrieval [SWS*00].

We introduce a novel concept to include image content in textual search interfaces by using state of the art methods. Furthermore, we investigate which image descriptors can be used to distinguish classes of visualizations and real world images. A first prototype to proof the concept is implemented that can measure the similarities between geospatial maps, line- and point-based visualizations and real world images. Finally, we reveal several directions for future work including overview visualization for exploration and automatic labeling of visualization techniques.

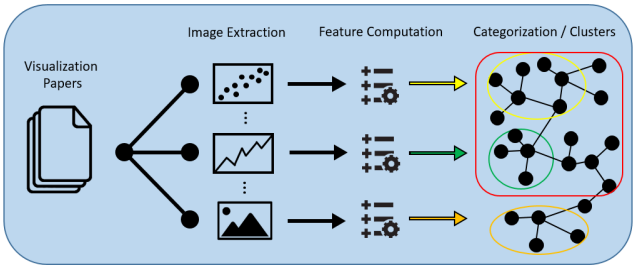

Figure 1: *Our approach to categorize images in documents. First, we extract the images from PDF documents. Then we compute features by using state of the art image descriptors, and finally, group similar visualization types via clustering. As a result, one could categorize scatter plots (yellow), time series (green), line and point-based visualizations (red) and real world images (orange).*

2. Concept

To compute image-based document similarity scores and apply image classification, images are extracted and image descriptors computed. Figure 1 illustrates our concept for the categorization of visualization types. Using image descriptors, images can be sorted, compared or clustered by their similarities. Image-based

DOI: 10.2312/eurp.20181116

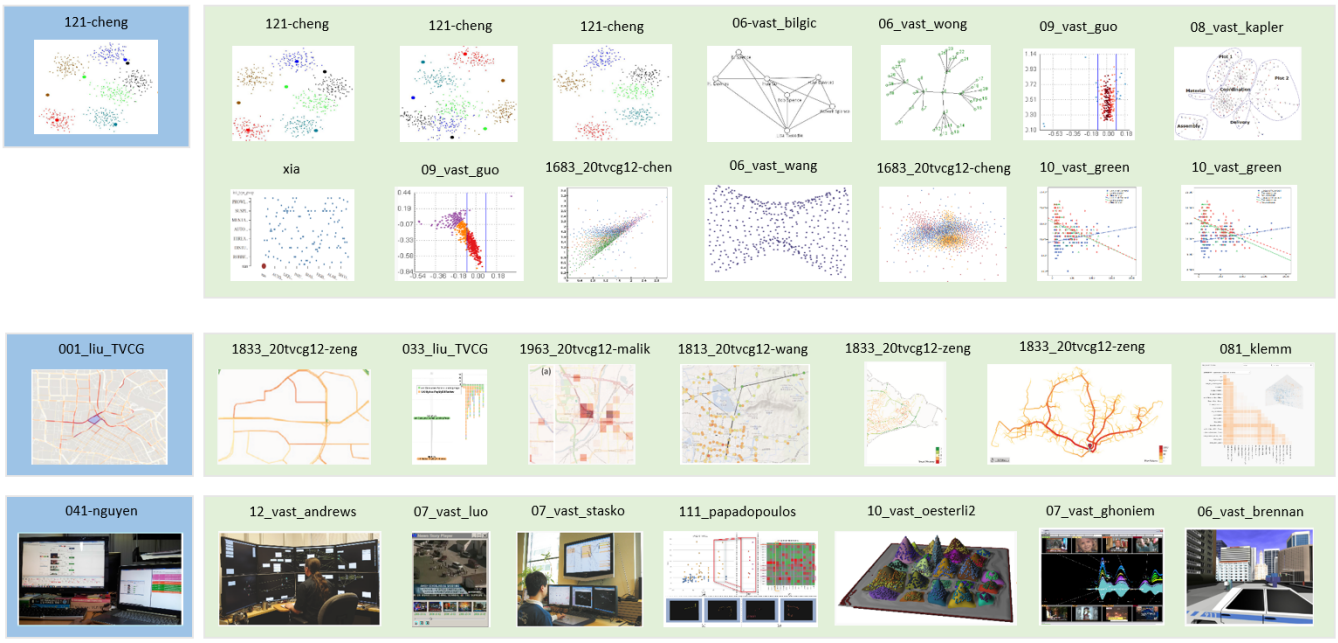

Figure 2: *Retrieval results of three different image types. Blue frame indicates queries, green frame search results and labels refer to filename. Our search system supports image retrieval to search for papers based on a query visualization, allowing to discover related techniques.*

descriptors have already been successfully used for scatter plot retrieval [SvLS12, SBS*14] and may be capable to classify other visualization types too. To this end, hierarchical clustering algorithms can be used to structure similar visualization types. It is desirable if we could automatically distinguish between different classes of visualizations such as, for high dimensional-, textual-, geospatial-data, abstract representations and real world images. Moreover, one could generate a dendrogram or graph of similar images to extend document exploration. The overall structure of image contents could be validated by including analysis on other meta-data such as labels, caption and keywords.

3. First Results

To achieve first results, we implemented a prototype including well-known information retrieval techniques (e.g., keyword search and tf-idf) and a Query-by-Example interface for image retrieval. We utilized the Apache PDFbox [pdf] to extract text and images from the PDF files. For computing feature vectors the Fuzzy Color and Texture Histogram descriptor (FCTH) from the open source visual information retrieval library Lire [LM13, lir] is used. The similarity of images is computed by the Tanimoto Coefficient (Jaccard coefficient).In our preliminary experiments, we used the VISPUB data set by Isenberg et al. [IHK*17]. We focus on the set of VAST documents from 2006-2017 and extract over 6.000 images for testing.

A primary goal was to identify which descriptors are suitable to distinguish different types of visualizations, as well as separating visualizations from real-word images. Figure 2 shows example image retrieval results by using a scatter plot, map visualization and a real world image (monitoring experiment) as query. One can see that the FCTH descriptor works quite well to distinguish all three classes. All queries returned reasonable results from different papers over the years. Of particular interest is to see that the scatter plot returns only point and line-based visualization techniques. The real world example shows the most false positives, although one may conceive some degree of similarity based on color and texture. We exemplify this approach by using papers from the Vis community but expect this concept to transfer to other domains as well.

4. Conclusion & Future Work

We introduced a novel approach to identify visualization classes in images and showed first image retrieval results of our initial prototype. The presented work in progress provides a solid base that we want to extend in the future in several directions. By integrating advanced text analysis methods automatic labeling of visualization could be further improved. In the future, we plan to develop an active learning model that should learn the essential features of one visualization (e.g., large number of edge crossings for parallel coordinates; a wider range of colors for real world images). It will be interesting to compare which kinds of image descriptors will work best on which kinds of visualizations. Possibly, new descriptors can be engineered to support specific visualization retrieval tasks. Finally, inspired by the *Phylogenetic Trees* of Li et al. [LCG*15], we intend to develop a visual representation of image and paper similarity relationships, which could help in defining new taxonomies of visualization techniques.

Acknowledgment

We thank Petra Isenberg and colleagues for providing us their VisPubData collection for experimentation.

References

[CGS*11] CAO N., GOTZ D., SUN J., LIN Y. R., QU H.: Solarmap: Multifaceted visual analytics for topic exploration. In *2011 IEEE 11th International Conference on Data Mining* (Dec 2011), pp. 101–110. doi:10.1109/ICDM.2011.135. 1

[IHK*17] ISENBERG P., HEIMERL F., KOCH S., ISENBERG T., XU P., STOLPER C. D., SEDLMAIR M. M., CHEN J., MÖLLER T., STASKO J.: vispubdata.org: A Metadata Collection about IEEE Visualization (VIS) Publications. *IEEE Transactions on Visualization and Computer Graphics 23* (2017). URL: https://hal.inria.fr/hal-01376597, doi:10.1109/TVCG.2016.2615308. 2

[LCG*15] LI S., CROUSER R. J., GRIFFIN G., GRAMAZIO C., SCHULZ H., CHILDS H., CHANG R.: Exploring hierarchical visualization designs using phylogenetic trees. In *Visualization and Data Analysis 2015, San Francisco, CA, USA, February* (2015), p. 939709. URL: https://doi.org/10.1117/12.2078857, doi:10.1117/12.2078857. 2

[lir] Lire: Lucene image retrieval. http://www.lire-project.net/. Accessed: 2018-04-27. 2

[LM13] LUX M., MARQUES O.: *Visual Information Retrieval Using Java and LIRE*, 1st ed. Morgan & Claypool Publishers, 2013. 2

[LZP*12] LIU S., ZHOU M. X., PAN S., SONG Y., QIAN W., CAI W., LIAN X.: Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Trans. Intell. Syst. Technol. 3*, 2 (Feb. 2012), 25:1–25:28. URL: http://doi.acm.org/10.1145/2089094.2089101, doi:10.1145/2089094.2089101. 1

[pdf] The apache software foundation - apache pdfbox. https://pdfbox.apache.org/. Accessed: 2018-04-27. 2

[SBS*14] SHAO L., BEHRISCH M., SCHRECK T., LANDESBERGER T. V., SCHERER M., BREMM S., KEIM D. A.: Guided sketching for visual search and exploration in large scatter plot spaces. In *EuroVis 2014 : the Eurographics Conference on Visualization ; 9-13 June 2014, Swansea, Wales, UK; EuroVA 2014* (2014), Pohl M., (Ed.), Eurographics Association, pp. 19–23. doi:10.2312/eurova.20141140. 2

[SOR*09] STROBELT H., OELKE D., ROHRDANTZ C., STOFFEL A., KEIM D. A., DEUSSEN O.: Document cards: A top trumps visualization for documents. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (Nov 2009), 1145–1152. doi:10.1109/TVCG.2009.139. 1

[SvLS12] SCHERER M., VON LANDESBERGER T., SCHRECK T.: A benchmark for content-based retrieval in bivariate data collections. In *Proceedings of the Second International Conference on Theory and Practice of Digital Libraries* (Berlin, Heidelberg, 2012), TPDL'12, Springer-Verlag, pp. 286–297. URL: http://dx.doi.org/10.1007/978-3-642-33290-6_31, doi:10.1007/978-3-642-33290-6_31. 2

[SWS*00] SMEULDERS A. W. M., WORRING M., SANTINI S., GUPTA A., JAIN R.: Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 12 (Dec 2000), 1349–1380. doi:10.1109/34.895972. 1