IMAGE REUSE DETECTION IN LARGE-SCALE DOCUMENT SCIENTIFIC COLLECTION

Oleg Bakhteev[1], Yury Chekhovich[1], Evgeny Finogeev[1], Tatiana Gorlenko[1], Mariam Kaprielova[1], Aleksandr Kildyakov[1], Aleksandr Ogaltsov[1]

[1]*The Antiplagiat Company, Moscow, Russia*

Abstract

In this report, we consider the problem of identification of image reuse cases in collections of scientific documents by means of an automatic image reuse detection system.

The problem's relevance is due to the presence of precedents of reusing images from other sources in the field of medicine and biology. Thus, in (Bik et al., 2016), it is shown that up to 4 % of reused images are found in scientific articles on biology and medicine.

In the latest decade, the problem of identification of image reuse has already been addressed in several works (Srivastava et al., 2015; Akshay et al., 2019; Meuschke et al., 2018). The rapid development of deep learning methods of image processing made it possible to create automatic searching systems of similar images in collections (Wang et al., 2014). Those systems can be adapted to the problem stated in this report. In (Srivastava et al., 2015; Akshay et al., 2019), authors apply classical computer vision methods to image reuse search. Those methods include image hashing algorithms (Tang et al., 2012; Yang et al., 2006) and keypoint detection by different algorithms (Lowe, 2004; Bay et al., 2006). Nevertheless, those approaches were tested on collections that consist of several thousands of images, while a collection retrieved from academic works can contain several millions of images. We analyzed (Srivastava et al., 2015) and recreated the experiment on a extensive data collection. This experiment showed low recall of the approach.

We developed a solution aimed to find image reuse in collections that contain several millions of images. It includes both classical computer vision algorithms and deep learning methods of image processing.

The technology of image reuse detection developed by us consists of four stages. We consider one of the images in the collection as the source of reuse. Different types of transformations could possibly be applied to the source (scaling, compression, rotation, reflection, greyscaling, channel selection, etc.)

The first step is to extract all the images from the document. Each page of the document is a separate image in high resolution. To get all the images from each page, we process every page of the document using the methods of classical computer vision, which highlight the images on the page. We do not use image extraction algorithms and libraries straightforward in order to avoid any influence of the way the document was generated.

At the second stage, charts, diagrams, schematics are excluded from images of the document. We do this to avoid a large number of false positives since diagrams will be easily recognized as similar to any incoming charts because the structure of images of this type are often very similar. Images that remain after the second stage are considered suitable for searching. In future work we plan to develop an independent solution for processing schematics and charts.

The third step is to search for candidates in the index of scientific documents. At this stage we form a fixed set of candidates from the collection for every suitable image. The special feature of this stage is the necessity to search in

the index. It is obvious that we can not compare incoming image with each object from the collection and perform the search for a reasonable time.

The fourth step is to match candidates with the right image accurately. To perform this stage, we use a Siamese neural network (Melekhov et al., 2016). We calculate similarity function between the matching image and each candidate to compare the candidates. Based on the values of the similarity function, it is determined both whether a given matching image is reused or not and the original source of the image.

We held an experiment in order to find cases of image reuse in articles from the list of open access journals DOAJ (DOAJ, 2022) using our solution. Our aim was to verify a hypothesis about the presence of multiple cases of image reuse. We also analyzed detected cases of image reuse and specified the nature of those cases.

We indexed 1,970,703 DOAJ articles and formed a collection of 6,081,847 images. Then we submitted each image as a request and checked throughout the collection. As a result, we found cases of reuse. All the results were analyzed by assessors and divided into three groups: supposedly incorrect reuse, correct reuse, reuse of images, published by the same author in another source. The results of the analysis are represented in the report.

This work was supported by FASIE (FASIE, 2022) project 63449.

Preliminary materials for this paper were published in the proceedings of the 20th Conference Mathematical Methods of Pattern Recognition (Bakhteev et al., 2021).

References

Akshay S., Chaitanya, B. N., & Rishabh, K. (2019). Image Plagiarism Detection using Compressed Images. *International Journal of Innovative Technology and Exploring Engineering*, *8*, 1423-1426.

Bakhteev O., Chekhovich Y., Finogeev E., Gorlenko T., Kaprielova M., Kildyakov A., & Ogaltsov A. (2021) Image reuse detection in large-scale document scientific collection. *Mathematical Methods for Pattern Recognition: Book of abstract of the 20th Russian National Conference with International Participation, Moscow, 2021,* 218-219.

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding*, *110*(3), 346–359. https://doi.org/10.1016/j.cviu.2007.09.014

Bik, E. M., Casadevall, A., & Fang, F. C. (2016). The Prevalence of Inappropriate Image Duplication in Biomedical Research Publications. *MBio, 7*(3). https://doi.org/10.1128/mBio.00809-16

Directory of Open Access Journals. DOAJ. (2022). https://doaj.org/

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints, cascade filtering approach. *International Journal of Computer Vision*, *60*, 91 – 110.

Melekhov, I., Kannala, J., & Rahtu, E. (2016). Siamese network features for image matching. *Proceedings of the 23rd international conference on pattern recognition*, Cancun, 378-383.

Meuschke, N., Gondek, C., Seebacher, D., Breitinger, C., Keim, D., & Gipp, B. (2018). An Adaptive Image-based Plagiarism Detection Approach. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, USA, 131–140. https://doi.org/10.1145/3197026.3197042

Srivastava, S., Mukherjee, P., & Lall, B. (2015). imPlag: Detecting image plagiarism using hierarchical near duplicate retrieval. *Annual IEEE India Conference (INDICON)*, India, 1-6. https://doi.ieeecomputersociety.org/10.1109/INDICON.2015.7443541

Tang, Z., Dai, Y., & Zhang, X. (2012). Perceptual hashing for color images using invariant moments. *Applied Mathematics and Information Sciences, 6.* 643-650.

The Foundation for the Promotion of Innovation. FASIE. (2022). https://fasie.ru/

Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., & Wu, Y. (2014). Learning fine-grained image similarity with deep ranking. *Proceedings of the IEEE conference on computer vision and pattern recognition,* USA, 1386-1393.

Yang, B., Gu, F., & Niu, X. (2006). Block Mean Value Based Image Perceptual Hashing. *International Conference on Intelligent Information Hiding and Multimedia*, USA, 167-172. https://doi.ieeecomputersociety.org/10.1109/IIH-MSP.2006.66