Finding duplicate images in biology papers

Markus Zlabinger
Vienna University of Technology, Austria
Favoritenstrasse 9-11/188, 1040 Vienna
markus.zlabinger@tuwien.ac.at

Allan Hanbury
Vienna University of Technology, Austria
Favoritenstrasse 9-11/188, 1040 Vienna
hanbury@ifs.tuwien.ac.at

ABSTRACT

Duplicated images in biology papers are a possible indicator for plagiarism or data fabrication. A manual detection of such duplicates can be time consuming or even infeasible for huge image collections. In this paper, a semi-automatic duplicate detection approach is proposed. The approach can be used for the detection of duplicates that cover only a fraction of the full image, are transformed (e.g. rotation), occur between images or within single images (i.e. single-image-duplicates). In the proposed approach, single-image-duplicates are detected between sub-images (i.e. sub-figures) based on a connected component approach and duplicates between images are detected via the min-hashing technique. The approach was evaluated on 1.7 million images extracted from biology papers. By application of various filtering methods to remove false positive detections, only a small amount of manual effort was necessary to find 3041 potentially serious duplicates in so far non-retracted papers.

CCS Concepts

•**Information systems → Information retrieval;**

Keywords

near-duplicate image detection; image retrieval

1. INTRODUCTION

Duplicated images in scientific publications are a possible indicator for plagiarism or data fabrication. The revelation of one of those practices causes the retraction of the affected work. Since the time of retraction can be years after publication, other authors may have already used the retracted work for their own research. Therefore, it is important to detect such duplicates as soon as possible.

Biology papers are particularly affected by inapropriate image duplication since they often contain illustrations of cells, Blots or other microscope images. Those image types are a favoured target for reuse or modification to pretend new results. In a recently published study [1], 20,621 biology papers were manually screened for image duplications. The results showed that 3.8% (1 out of 26) of the examined papers contained problematic duplicates, whereby half of them were considered as the result of intentional misconduct.

To find duplicated images in biology papers, manual screening can be conducted, as done in the mentioned study in [1]. However, two things need to be considered: First, manual screening can be time consuming for huge collections that contain million of images and second, duplicates between images are only detectable within small image groups because the number of possible pairs grows with the number of images N as $\binom{N}{2}$.

In this paper, a semi-automatic approach is proposed, which can be used to find potentially serious duplicates in biology papers. The main contributions of the approach are the following: First, it can be used for the detection of duplicated areas between images, but also for the detection of duplicated areas that occur within a single image. Note that for the rest of this paper, a duplicate detected between two images is called *double-image-duplicate* (DID) and a duplicate that is detected within one single image a *single-image-duplicate* (SID). Second, the approach is applicable on large image collections and works even if transformations (e.g. rotation) were applied and finally, detections that are usually irrelevant in regards to plagiarism or data fabrication (e.g. duplicates based on text labels, diagram axes, ...) are discarded by the application of various filtering methods. Especially, the duplicate detection within single images and the removal of irrelevant detections is not addressed by current literature.

2. BACKGROUND AND RELATED WORK

To find DIDs an efficient pairing method is necessary. The state-of-the-art methods (e.g. [3, 2]) are based on the Bag-of-Words model (BoW) and the min-hashing technique. In the BoW model, multi-dimensional descriptors (e.g. SIFT [4]) are mapped to single integer values, the visual words. Visual words are used as input for the standard min-hashing algorithm [3], which can be summarized as follows: Random hash-functions are used to select visual words randomly, i.e.

SAC 2017,April 03 - 07, 2017, Marrakech, Morocco
Copyright 2017 ACM 978-1-4503-4486-9/17/04...$15.00
http://dx.doi.org/10.1145/3019612.3019875

the word with the smallest value (i.e. the minimum hash) is selected. Then, the selected words are bundled and afterwards, the bundles, which are also called sketches, are put into a hash-table. Colliding images within the hash-table (i.e. identical sketch) are paired.

Chum et al. [2] improved the standard min-hashing algorithm by encoding geometric information into the sketches and called the method *Geometric Min-Hashing* (GmH). In the GmH algorithm, only the first sketch value is selected randomly from the visual words and subsequently selected values must be within a minimum distance and additionally, must have a similar scale to the initially selected word. The distance and scale information is based on the underlying keypoints of the visual words, e.g. *Difference of Gaussians* (DoG). The rest of GmH is equivalent as in the standard min-hashing approach (e.g. hash-table insertion, ...).

In this paper's approach, the GmH algorithm was used for the detection of DIDs because of its properties [2]: First, it can be used for both the discovery of duplicates (i.e. initial detection of duplicated images within a database) and the retrieval of duplicates (i.e. querying of a new image against an existing database). Second, duplicated areas that cover only a small fraction of the full image (coverage of about 0.28% [2]) can be detected and finally, the algorithm works on large image collections.

The configuration of the GmH algorithm (e.g. number of sketches) was selected as in the original paper [2]. For the mapping of visual words, a one million word vocabulary was used (considered as a reasonable choice in [3, 2]). Note that also for the other methods that make use of features in this paper (e.g. descriptor matching), SIFT descriptors and DoG keypoints were selected, because [4]: First, they are invariant to rotation, various affine distortions, addition of noise and change in illumination and second, the features are highly distinctive (i.e. matched SIFT descriptors are a strong indicator for duplicated areas).

3. DUPLICATE DETECTION METHOD

The proposed approach consists of three steps: First, diagrams are removed from all images in a collection (described in Section 3.1). Second, within the remaining images after diagram removal, initial SID (Section 3.2) and DID pairs (Section 3.3) are detected. Finally, the detected pairs are filtered from false positives (Section 3.4).

3.1 Diagram Removal

Diagrams are an essential part of scientific publications and occur very frequently. Their frequency and similarity (e.g. axis labels, bars) would lead to a high number of false positive detections, which makes manual evaluation infeasible. Therefore, diagrams are removed from each input image as follows: First, connected components are extracted. Second, a rectangular bounding box if fitted to each connected component, whereby each bounding box represents a sub-image (i.e. sub-figure). Then, bounding boxes that enclose an area smaller than 1000 pixels (see Section 4) are discarded (e.g. to discard isolated characters, single black dots, ...). From the remaining sub-images, visual words are extracted and finally, the visual words are classified via a *Support Vector Machine* (SVM) classifier as diagram or non-diagram.

The removal of diagrams has three benefits: First, duplicates are not detected between diagrams (higher precision). Second, GmH sketches are not extracted from diagram areas, which increases the number of sketches from relevant image areas (higher recall). Third, images containing exclusively diagrams can be ignored (reduced resource requirements).

3.2 Single image pairing

To find SIDs, the sub-figure structure of scientific images is exploited as follows: First, sub-images are extracted (see Section 3.1). Then, SIFT descriptors are computed for each image. Afterwards, each descriptor is assigned to a sub-image (based on its x and y coordinates). Finally, the sub-images are paired exhaustively and SIFT descriptors are matched between the two sub-images of each pair. Pairs with at least three matched descriptors [4], after removal of ambiguous matches (described in Section 3.4), are the initially detected SID pairs. The further applied processing steps for each pair are described in Section 3.4.

3.3 Double image pairing

For the detection of DIDs, the GmH algorithm is used (see Section 2). However, before GmH sketches are extracted, 10% (see Section 4) of the most frequently occurring visual words (stop words) are removed. Those words usually describe irrelevant (in regards to plagiarism or data fabrication) image areas (e.g. the label areas in Figure 1) and can be removed to prevent false positive detections between them. The remaining visual words are used as input for the GmH algorithm. Afterwards, images with identical GmH sketches are paired. Finally, for each pair, the set similarity [3] is calculated and pairs with less than $sim_{min} = 0.08$ (Section 4) are discarded. The set similarity between two sets of visual words S_1 and S_2 is defined as the words in their intersection, i.e. $|S_1 \cap S_2|$, divided by the number of words in their union, i.e. $|S_1 \cup S_2|$. Pairs that are not discarded are the initially detected DID pairs.

3.4 Descriptor matching & label filter

In the previous two sections, the pairing of SIDs and DIDs was described. Now, false positives are filtered by application of two filtering methods. The first filtering method is the discarding of pairs with less than $M_{min} = 14$ (see Section 4) SIFT descriptor matches. Initial matches for this filtering method are detected between sub-images for SID pairs and between images for DID pairs. As distance measure the Euclidean distance is used. Ambiguous matches are discarded by application of the *Ratio test* [4] and outlier removal via the RANSAC algorithm. Pairs with less than 14 matches, after removal of ambiguous matches, are discarded.

In a final filtering step (called *label filter* and first introduced in this paper), duplicates that are detected based on labels (see examples in Figure 1) are discarded as follows: First, matched keypoints in both images (or sub-images for SID pairs) of a pair are clustered if they are located within a distance of $d_m = 75$ pixels (see Section 4). The idea is to cluster keypoints that describe the same duplicated area, for example, if there are SIFT matches in the top left (e.g. a sub-

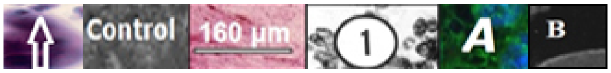

Figure 1: Labels that may lead to false positives

figure label B) in both images/sub-images and a matched magnification factor label (e.g. 200μ) in the bottom right, four clusters should be detected (two in each image/sub-image). Second, a rectangular bounding box is fitted to each cluster and boxes that enclose an area with less than $lf_{min} = 2000$ (see Section 4) pixels are discarded. Note that empirical testing showed (see Section 4) that true positives usually have a duplicated area larger than 2000 pixels and are therefore preserved.

4. RESULTS & DISCUSSION

Dataset. To evaluate the approach, 1,721,612 images from *PubMed Central*[1] publications were extracted. Additionally to these images, 19 DID pairs (i.e. 38 images) and 10 SID pairs (i.e. 10 images) from already retracted papers (collected from *Retraction Watch*[2]) were added as needles (known duplicates). The collected needles consisted of various biological image types (blood cells, cancer cells, Western Blots[3], ...), fully duplicated images and a variety of partly duplicated images (e.g. coverage 1.5%, 9.4%, ...). The goal was to preserve as many needles as possible and at the same time, find potentially serious duplicates in the 1.7 million images. Note that the described evaluation approach was chosen since there does not exist a dataset with known ground truth in regards to plagiarism and data fabrication detection in biological images. In particular, the removal of irrelevant duplicates and detection of SIDs is not considered by existing datasets.

Diagram removal. From a sub-set of the 1.7 million images, 1228 diagram and 973 non-diagram images were manually extracted and used for training and testing of the SVM classifier. The best result, with 0.953 recall and 0.879 precision, was achieved using a 1500 word vocabulary (tested from 250 to 16,000) in combination with a *Radial Basis Function* (RBF) trained classifier (tested linear and RBF). When applied on the PubMed Central dataset, only 803,765 of the initially 1,721,612 images remained (i.e. the discarded images contained diagrams exclusively). On the other hand, the duplicated areas of the needle images were preserved.

Parameter selection. For the approach, it was necessary to select several parameters (e.g. sim_{min}, M_{min}, stop word removal, ...). To find the optimal parameter values in regards to false positive rate (FPR) and true positive rate (TPR), empirical testing was conducted, i.e. for each parameter, the critical boundary values were tested and the best performing value selected. The TPRs were calculated based on the needle images (i.e. 38 DID and 10 SID images) and the FPRs on the PubMed Central images (i.e. the 1.7 million images). The selected optimal parameters at each filtering step are included in Table 1.

Table 1: Results on the PubMed Central dataset

After step	DID pairs	SID pairs
GmH/sub-image pairing	29 362 174	339 224
Set sim. ($sim_{min} = 0.008$)	6 070 823	-
Min. matches ($M_{min} = 14$)	304 875	85 270
Label filter ($lf_{min} = 2000$)	245 155	68 703

Manual evaluation. After application of the duplicate detection algorithm with tuned parameters on the PubMed Central dataset (i.e. the 803,765 images after diagram removal), 245,155 DIDs and 68,703 SIDs were found (Table 1). All of these pairs contained duplicated (or at least very similar) areas, however, some of them were not relevant in regards to plagiarism or data fabrication, e.g. detections based on progression images (e.g. infection development) or magnified images (e.g. two images show the same object but once with 10x magnification and once with 20x). In a final manual screening, irrelevant detections were discarded. In the end, 1674 SIDs and 1367 DIDs remained, which were considered as potentially serious. The screening was conducted by a single person and took 7 hours for the DIDs and 24 hours for the SIDs, however, note that a large number of the 245,155 DIDs was based on image clusters, e.g. 422 images contained an identical map of Africa and resulted in the detection of $\binom{422}{2} = 88,831$ pairs. By blacklisting of the two most frequently occurring image types (maps, generic brain images), only 18,821 DID pairs remained, i.e. only 18,821 DID pairs instead of 245,155 were manually screened.

5. CONCLUSION

In this paper an image duplicate detection approach, to combat plagiarism and data fabrication, was proposed. It was shown that the approach can be used to find potentially serious duplicates, however, for a final judgement biology expertise on the duplicated images and their textual description is required. There is much room for future research (e.g. different visual features, additional filter methods), but to make different approaches comparable, it is first necessary to create a dataset with known ground truth.

6. REFERENCES

[1] E. M. Bik, A. Casadevall, and F. C. Fang. The prevalence of inappropriate image duplication in biomedical research publications. *mBio*, 7(3):e00809–16, 2016.

[2] O. Chum, M. Perd'och, and J. Matas. Geometric min-hashing: Finding a (thick) needle in a haystack. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 17–24. IEEE, 2009.

[3] O. Chum, J. Philbin, M. Isard, and A. Zisserman. Scalable near identical image and shot detection. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, pages 549–556. ACM, 2007.

[4] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[1] http://www.ncbi.nlm.nih.gov/pmc/

[2] http://retractionwatch.com/

[3] Used to detect and analyse proteins in samples