Finding Groups of Duplicate Images in Very Large Datasets

Winn Voravuthikunchai
winn.voravuthikunchai@unicaen.fr
Bruno Crémilleux
bruno.cremilleux@unicaen.fr
Frédéric Jurie
frederic.jurie@unicaen.fr

GREYC — CNRS UMR 6072,
University of Caen Basse-Normandie,
Caen, France

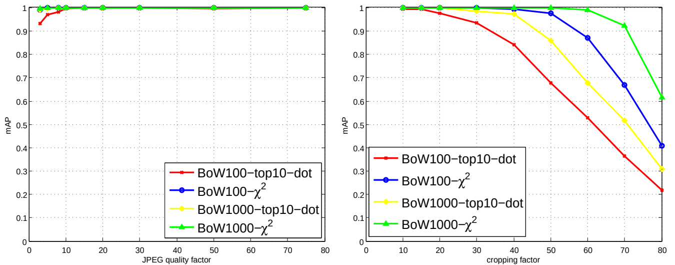
This paper addresses the problem of detecting groups of duplicates in large-scale unstructured image datasets such as the Internet. Leveraging the recent progress in data mining, we propose an efficient approach based on the search of *closed patterns*. Moreover, we present a novel way to encode the images based on bag-of-words vectors inspired by the text processing literature, that can be transformed into data mining transactions. Unlike other existing approaches, our method can scale gracefully to larger datasets as it has linear time and space (memory) complexities.

To encode the images as data mining transactions, we represent images by lists of their most top K informative visual words, using tf-idf weighting (term frequency-inverse document frequency). Tf-idf has been successful for normalizing BoW in vision tasks [2]. After representing images as transactions of items, we extract *all* closed itemsets whose length is greater than a given threshold (denoted *minlength*) as the length reflexes the similarity among the images containing the itemset. The minimum frequency (denoted *minfr*) support has to be set to 2 as two images can form a group of duplicates. Our mining strategy is based on LCM [3], which is one of the most efficient algorithms for mining frequent closed itemsets.

In our experiments, we first validate our proposed image binary representation in an image search scenario using the **Copydays dataset** [1]. Each of the 157 original image is used as a query to retrieve its corresponding attack image which is mixed in a set of 10,000 images. The average precision is computed and the mean is reported for all of the 157 queries. We compare the BoW based binary representation (with $K = 10$ and dot product similarity) to a standard BoW representation (with χ^2 distance) using two vocabularies size i.e. 100 and 1,000. We do the comparison by using the JPEG attacks (from JPEG3 to JPEG75) shown in Figure 1a and by using the cropping attacks (from 10% to 80%) shown in Figure 1b. In conclusion, this experiment demonstrate that this representation is sufficient for detecting near duplicate images, while being very compact (each image is encoded by ~13 bytes only).Furthermore, as this representation is made of lists of items, it can be used efficiently for finding frequent closed patterns.

The following experiment validates the mining algorithm proposed for discovering groups of duplicate images. We use again the **Copydays dataset**, but in a different way: in these experiments, we put together the 157 original images, their corresponding attacked images, and 1,000,000 artificial image descriptors. In the ideal case, the algorithm should correctly discover the 157 groups of duplicates. The performance is evaluated by using the mean F-score which is equal to one only if the system outputs exactly 157 groups containing the original image and its transformations. Figure 2 shows the F-Score as a function of *minlength*, for representations made from 100 and 1,000 visual words dictionaries. *minlength* = 7 and 1,000 visual words dictionary give optimal results. We can see that for the light attacks, the groups of images are perfectly detected. Even for the strongest attacks the results are still very good.

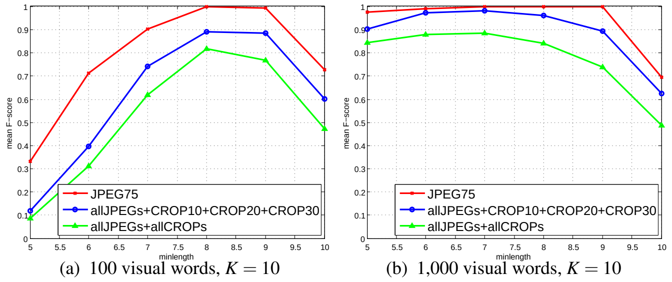
Then we perform our method to detect groups of duplicates on our **One million random web images database**. We show that the computation time and the memory usage scales linearly with the size of the dataset in Figure 3b and Figure 3c. We are able to obtained more than 80 thousands groups of duplicates in less than 3 minutes. Figure 4 shows some of these groups. Beside computational efficiency, these results demonstrate the robustness against compression, scaling, slight crops, rotation, insertion/removal of small elements, brightness/contrast changes.



(a) JPEG attacks from JPEG3 (very low quality) to JPEG75 (typical web quality).

(b) Cropping attacks from 10% to 80% of the image surface.

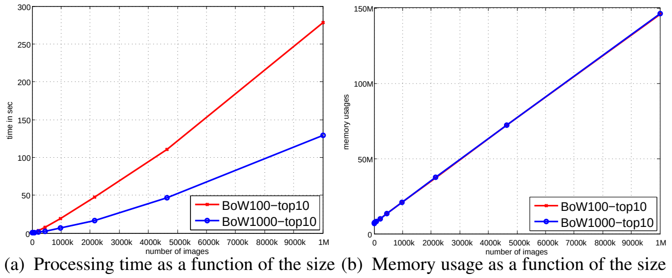
Figure 1: Image retrieval experiments: performance of the proposed representation and of the baseline representation, for two vocabularies (100 and 1,000 visual words).



(a) 100 visual words, $K = 10$

(b) 1,000 visual words, $K = 10$

Figure 2: Mean F-score as a function of *minlength*



(a) Processing time as a function of the size of the database

(b) Memory usage as a function of the size of the database.

Figure 3: Computation time and memory usage as a function of the number of images

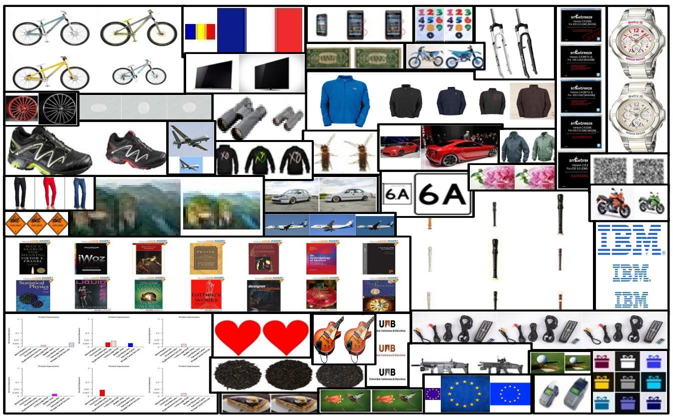


Figure 4: Some of the groups of duplicate/similar images found on the **One million random web images database**

[1] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid. Evaluation of gist descriptors for web-scale image search. In *CIVR*, pages 19:1–19:8, 2009.

[2] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[3] T. Uno, T. Asai, Y. Uchida, and H. Arimura. An efficient algorithm for enumerating closed patterns in transaction databases. In *proceedings of Discovery Science (DS'04)*, volume 3245 of *LNAI*, pages 16–31, Padova, Italy, 2004. Springer.