Identification of MIR-Flickr Near-Duplicate Images
a benchmark collection for near-duplicate detection

Richard Connor[1], Stewart MacKenzie-Leigh[1], Franco Alberto Cardillo[2] and Robert Moss[1]
[1] *Department of Computer and Information Sciences, University of Strathclyde, Glasgow, Scotland*
[2]*Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy*
richard.connor@strath.ac.uk, stewartml@gmail.com, franco.alberto.cardillo@isti.cnr.it, robertgmoss@gmail.com

Keywords: near-duplicate image detection, benchmark, image similarity function, forensic image detection

Abstract: There are many contexts where the automated detection of near-duplicate images is important, for example the detection of copyright infringement or images of child abuse. There are many published methods for the detection of similar and near-duplicate images; however it is still uncommon for methods to be objectively compared with each other, probably because of a lack of any good framework in which to do so. Published sets of near-duplicate images exist, but are typically small, specialist, or generated. Here, we give a new test set based on a large, serendipitously selected collection of high quality images. Having observed that the MIR-Flickr 1M image set contains a significant number of near-duplicate images, we have discovered the majority of these. We disclose a set of 1,958 near-duplicate clusters from within the set, and show that this is very likely to contain almost all of the near-duplicate pairs that exist. The main contribution of this publication is the identification of these images, which may then be used by other authors to make comparisons as they see fit. In particular however, near-duplicate classification functions may now be accurately tested for sensitivity and specificity over a general collection of images.

1 INTRODUCTION

Our primary interest is in the quantitative comparison of different similarity functions for the detection of near-duplicate images. Of particular interest to us is a "batch mode" of processing, necessary in forensic image detection, where two very large collections (e.g. each upwards of 10^6 images) require to be tested against each other with the intent of determining the near-duplicate intersection. In such cases, a high value for specificity of the similarity function is of paramount importance, to avoid the generation of a huge number of false positive matches. The similarity function is necessarily used with a threshold limit to give a classification function, rather than as in a more common ranking scenario. This shifts the balance of importance of the relative values of sensitivity and specificity (effectively, recall and precision) as any significant degree of imprecision will be completely unacceptable, and will always be traded for a loss of recall.

To measure these values for a given similarity function requires very large sets of benchmark images, with a known ground truth of near-duplicates. Furthermore, there should be no bias as to the type of images in the collection, nor the method with which the near-duplicates have been formed. These requirements seem to be mutually incompatible.

While performing analysis over the MIR-Flickr collection of one million images (Huiskes and Lew, 2008), we observed a significant number of near-duplicates. This in turn led us to realise that, if we could identify all of these, we would have a single collection of one million images which could be tested against itself, and would effectively have these desired properties.

In the absence of a perfect near-duplicate finder, it is of course impossible to find all the near-duplicate image pairs within the collection. However, using a number of different near-duplicate finders, we have found around 2,000 pairs of images conforming to an objective definition of near-duplicate. Using techniques from population statistics, we are able to confirm that these constitute almost all the pairs that exist within the collection.

The main contribution of this paper is the publication (Connor, 2015)[1] of our analysis of the image set, which can be used to rate different near-duplicate

[1]Available from
www.mir-flickr-near-duplicates.appspot.com

finding functions against each other, and to give accurate absolute values for sensitivity and specificity.

2 RELATED WORK

Since the main contribution of the work presented here is a new dataset that can be used in the context of near duplicate (ND) identification, in this section we present a brief review of existing datasets and of their usage in past work. Algorithms for the ND problem can be roughly classified into two categories according to the type of the input data they were created for, whether video files or image collections (Kim et al., 2010). Methods in the first category attempts to detect near duplicate keyframes (NDK) in video files. Such methods are mostly based on local visual features, such as SIFT, and are validated using standard benchmarking datasets, such as, for example, the TRECVID collection (Over, 2014).

For the second category of methods, whose aim is to return all the images in a collection that are duplicate or near-duplicate of a query image, the experimental context is not so well defined. In fact various authors state that they could not find any specific benchmark for the testing their novel approach: *"We do not have access to ground-truth data for our experiments, since we are not aware of any large public corpus in which near duplicate images have been annotated."* (Chum et al., 2007; Jinda-Apiraksa et al., 2013); *" Although the target application of this dataset is image retrieval, it was selected due to the lack of other appropriate datasets [...]"* (Vonikakis et al., 2014).

Many previous works simply use datasets created for multimedia information retrieval, such as the INRIA Holidays Dataset (Jegou et al., 2008). This dataset is composed of 1491 images, partitioned into 500 groups corresponding to 500 different scenes: the first image in each group is to be used as the query image and the remaining photos represent the correct result to be returned. However, there is no information about duplicate or near-duplicate images.

Some works use the dataset presented in (Nister and Stewenius, 2006), which is composed of 10,200 images in sets of 4 images of one object / scene. Even if the sets might be used to evaluate the performance of a near-duplicate finder, there is no information about how similar two sets might be and whether or not they should be considered duplicate or near-duplicate.

In (Jinda-Apiraksa et al., 2013) the authors give a dataset specifically built for near-duplicate image detecetion. The dataset is composed of 701 photos taken during a trip. The photographies were not digitally manipulated and should represent a real photo gallery of a generic user. They include changes in scene, camera, and image as defined in (Jaimes et al., 2003). A well-established ground-truth is provided with the dataset based on the judgments provided by ten different individuals; however the set is rather small for general deductions to be made. To the best of our knowledge this is the only public domain dataset which has been built specifically for the task of near duplicate detection.

3 MIR-FLICKR

The MIR-Flickr image dataset (Huiskes and Lew, 2008; Huiskes et al., 2010) consists of one million "interesting" images downloaded from the website flickr.com through its public API. The "interestingness" of the images represents a score attributed by the flickr service by taking into account the comments and the clickthroughs on the images. For each image in the dataset, the authors provide the flickr tags, the exif metadata, plus global (edge histogram, homogeneous local texture, gist) and local visual descriptors (SURF). Since the 1M images included in the dataset were not selected with a specific task or set of criteria in mind, they should represent a good benchmark for evaluation near duplicate detection algorithms on large image datasets.

4 NEAR-DUPLICATE IMAGES

We have identified near-duplicate images in two categories, defined in (Foo et al., 2006) as *identical* and *non-identical* near duplicates (*IND* and *NIND* respectively). *IND* images are "derived from the same digital source after applying some transformations", and *NIND* images "share the same scenes and objects".

We interpret the notion of *transformation* to include any operation which has been performed using a standard image editor, with the intent of making cosmetic changes. While this definition is not completely objective, we have found it to be effective in that different humans seem to generally agree on the classification of images based on it. Any remaining inherent subjectivity is safeguarded by the publication of our classification based on this description.

For the purposes of benchmarking, we choose to primarily use the *IND* definition for the following reasons:

- it is (almost) objective
- such pairs are relatively common in the MIR Flickr set
- it is the most useful concept for forensic image detection
- we can be much more confident of identifying the vast majority of pairs within the set
- the resulting relation is, effectively, an equivalence relation, allowing the identification of some near-duplicate clusters containing more than two images

We have, however, also identified as many *NIND* pairs as possible and also provide these in our published data.

4.1 Pairs and Clusters

The identification of clusters, rather than pairs, is important. The largest *IND* cluster found contains 15 images, therefore giving 105 unique near-duplicate pairs. The presence of clusters of size greater than two is completely arbitrary, and it would be incorrect to allow it to influence the measurement of similarity functions which may, or may not, be particularly suited to the type of image in the cluster.

It also seems important not to discriminate against similarity measures detecting pairs of images which are visually very similar, but which do not meet the strict criteria of the definition. For this reason pairs of images were classified in three categories:

1. *IND* near-duplicates, as defined above
2. pairs of images which are strikingly visually similar, but are not *IND* as defined [2], and
3. pairs which do not meet either criteria

Figures 1 and 2 gives examples of the first two categories; all identified pairs in these categories are published at (Connor, 2015).

The *IND* relation in this context is reflexive, symmetric and transitive. Transitivity is not a property of near-duplication in general, but is a safe assumption for our target set and definition. As *IND* is thus an equivalence relation, it defines a partition over the image set. The set of near-duplicate clusters is defined as the set of all equivalence classes whose cardinality is two or greater.

[2]There include some pairs which do not strictly match the *NIND* definition, such as generated images

Figure 1: *IND* near-duplicate images (images 88518 and 90355)

Figure 2: "strikingly similar", but not near-duplicate (images 46271 and 47850)

5 METHODOLOGY

5.1 Characterisations and Metrics

To discover near-duplicate images within MIR-Flickr, a number of different distance metrics (Table 2) were applied to a number of different image characterisations (Table 1). The characterisations chosen represent global, rather than local, features, as these should be better near-duplicate detectors according to our definition.

Table 1: Image characterisations used

Eh	MPEG-7 Edge Histograms (Won et al., 2002)
Ht	MPEG-7 Heterogeneous Textures (Bober, 2001)
Cs	MPEG-7 Colour Structures (Bober, 2001)
pHash	Perceptual Hashing (Niu and Jiao, 2008)

Table 2: Distance Metrics used

Man	Manhattan (L_1) distance
Euc	Euclidean (L_2) distance
Cos	Cosine distance
Sed	Structural Entropic Distance
Ham	Hamming distance over bitmaps

By "Cosine Distance" we refer to the proper metric form[3]; "Structural Entropic Distance" refers to the distance metric defined in (Connor et al., 2011) and refined in (Connor and Moss, 2012). *Ht* and *Eh* data was taken from the MIR-Flickr site; *Cs* and *pHash* data was extracted by our own code according to the published specifications.

Hamming distance was applied to *pHash*, and the other distances were all applied to all the other characterisations, giving a total of 15 different distance functions.

5.2 Cluster Identification

The following method was used to produce the near-duplicate clustering:

1. The data set was first cleaned to remove images that were a perfect duplicate of another, defined as being the same size with the same pixel values at each location. 378 images were removed at this stage.
2. For each similarity function, a threshold-limited nearest-neighbour search was conducted for each image in the set. This requires potentially 10^{12} comparisons, which is infeasible for almost any cost of comparison, and the number of pairs elicited depended on various pragmatic cost factors. However we were able to extract a least a few thousand pairs for every function. These computations are still extremely compute-intensive, and metric search techniques (Chávez et al., 2001; Zezula et al., 2006) were used.
3. Each of the resulting image pairs was inspected by a member of the project team and judged to be in one of the three categories explained above.
4. The resulting set of positively identified *IND* pairs, from all metrics, was treated as a set of clusters of size 2, which were then rationalised by (repeatedly) amalgamating any clusters which had a common element.

At point of publication, this has resulted in the identification of 1,958 near-duplicate clusters within the set, containing a total of 4,071 images. The mean size of a cluster is 2.08. 543 pairs of "strikingly similar" images have been identified. The identities of all these images are given, along with views onto the images themselves, at(Connor, 2015).

6 ESTIMATE OF TRUE SIZE

The observation that the size of a population can be estimated from a number of independent, imperfect counts was first made by Laplace in the 18th Century. The context is that two independent detectors A, B detect a, b instances of a phenomenon respectively, and z instances are detected by both. The detectors have unknown yet consistent detection probabilities p_A, p_B. For a total (large) number of occurrences N, then $a \approx p_A N$ and $b \approx p_B N$. For the number detected by both, $z \approx p_A p_B N$. Therefore $N \approx \frac{ab}{z}$.

This observation was extended and refined by Chapman, an elegant description being given in (Pollock et al., 1990), as:

$$\hat{N} = \frac{(a+1)(b+1)}{(z+1)}$$

and an estimate of the variance of the outcome is also given:

$$\hat{V} = \frac{(a+1)(b+1)(a-z)(b-z)}{(z+1)^2(z+2)}$$

[3] the angle between the vectors rather than the complement of its cosine, which is not a proper metric

which allows confidence intervals to be assigned.

We have made three such estimations, through taking the three apparently most independent image characterisations, and using, for each characterisation, the metric which retrieved the most pairs. The results of this are shown in Table 3, suggesting a true value of a little less than 1,900.

In fact, from all characterisations and metrics tested, we have so far found a total of 1,958 clusters of images. While there is probably some interdependence among the methods we have used, which would have a tendency to reduce the derived estimates, there is certainly significant independence as evidenced by the relatively small intersection sizes. We therefore judge the value for the whole population to be somewhere very close to this value. The probability of the true collection size being greater than, for example, 2,000 seems to be very close to zero, allowing this figure to be used as a (conservative) basis for measuring true values of sensitivity and specificity.

7 SEMANTIC COMPARISON

The purpose of establishing the benchmark set is to allow a useful comparison of different near-duplicate detection functions, and we are now embarking upon a deeper study of these. However, we already of course have results for the functions used to construct the set. Here we show only simple results for the three functions used to construct the population estimate to give a flavour of one way the benchmark set can be used.

For each function graphs are shown of sensitivity, assuming a true collection size of 2,000 clusters, and positive predictive value (PPV, commonly known as precision in information retrieval) both measured against the threshold at which the function is applied. Each graph is plotted past the crossover points of these two values, in all cases incorporating at least the 2,000 nearest-neighbour pairs with the smallest distances.

While this is, at this point, a relatively shallow comparison, this is the first time, to the best of our knowledge, that any two such classification functions have been objectively compared with each other over a large collection. Even this gives a clear visual indication that the *Eh/Sed* classification function is significantly the best of those tested, and that this function could be used in "batch" mode with a threshold that will give almost no false positives, and find approximately half of the *IND* intersection between two large sets of images. This in itself gives a significant result in the domain of forensic image detection.

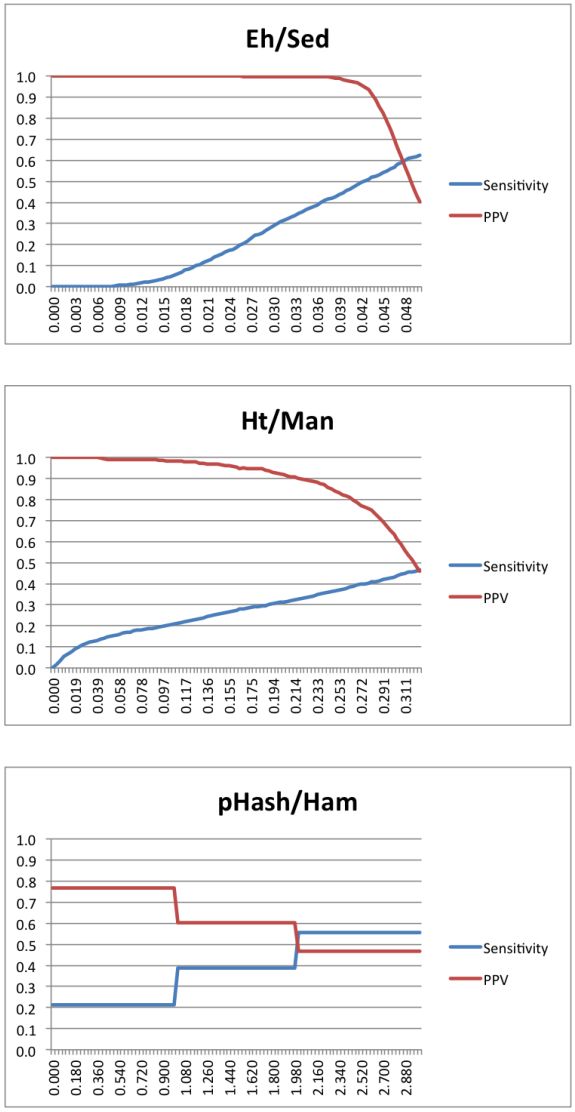

Figure 3: Semantic comparison of the best independent detection functions

One further point of more general interest is that, in many cases, little attention is paid to the distance metric used with any single characterisation. In particular, edge histograms are generally used with L_1 distance, and heterogeneous textures with L_2. It is noteworthy that in neither case is this the best metric, and in fact in the case of edge histograms, all of the other metrics tested significantly outperform Manhattan distance semantically. This outcome in itself highlights the importance of collections such as the one we have established, as it allows this type of measurement to be objectively performed, which is not possible with a small image collection.

Table 3: Population Estimates

Method 1	Method 2	a	b	z	$\hat{N}$	$\hat{V}$	98% CI
Eh/Sed	Ht/Man	1225	916	579	1938	1252	1868-2009
Eh/Sed	pHash/Ham	1225	1130	754	1837	570	1789-1884
pHash/Ham	Ht/Man	1130	916	560	1849	1190	1780-1918

8 CONCLUSIONS

We have identified and published a set of nearly 2,000 near-duplicate clusters which occur within the MIR-Flickr image collection of one million images. As both the collection and the near-duplicate subset have occurred through serendipitous processes, this makes a valuable test set for the semantic comparison of near-duplicate finding functions. While work using the test set is still at an early stage, we have already made some surprising discoveries in terms of the use of different metrics with well-known image characterisation functions.

The exhaustive search for near-duplicates within the set will of course never be finished: any updates will be gratefully received by the authors, and communicated onwards through our website.

ACKNOWLEDGEMENTS

We would like to acknowledge help and advice from Mark Huiskes of Leiden University and Kenneth Pollock of North Carolina State University, for sharing their knowledge about the MIR-Flickr collection and population statistics respectively. We would also like to thank Richard Martin and Karina Kubiak-Ossowska of the University of Strathclyde for help with access to the ARCHIE-WeSt HPC facilities necessary to achieve some of the analysis.

Franco Alberto Cardillo was supported by the National Research Council of Italy (CNR) for a Short-term Mobility Fellowship (STM), which funded a stay at the University of Strathclyde in Glasgow (UK) where part of this work was done.

REFERENCES

Bober, M. (2001). Mpeg-7 visual shape descriptors. *IEEE Transactions on circuits and systems for video technology*, 11(6):716–719.

Chávez, E., Navarro, G., Baeza-Yates, R., and Marroquín, J. L. (2001). Searching in metric spaces. *ACM Comput. Surv.*, 33(3):273–321.

Chum, O., Philbin, J., Isard, M., and Zisserman, A. (2007). Scalable near identical image and shot detection. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 549–556. ACM.

Connor, R. (2015). Mir-flickr near-duplicate data. `mir-flickr-near-duplicates.appspot.com`.

Connor, R. and Moss, R. (2012). A multivariate correlation distance for vector spaces. In Navarro, G. and Pestov, V., editors, *Similarity Search and Applications*, volume 7404 of *Lecture Notes in Computer Science*, pages 209–225. Springer Berlin Heidelberg.

Connor, R., Simeoni, F., Iakovos, M., and Moss, R. (2011). A bounded distance metric for comparing tree structure. *Inf. Syst.*, 36(4):748–764.

Foo, J., Sinha, R., and Zobel, J. (2006). Discovery of image versions in large collections. In Cham, T.-J., Cai, J., Dorai, C., Rajan, D., Chua, T.-S., and Chia, L.-T., editors, *Advances in Multimedia Modeling*, volume 4352 of *Lecture Notes in Computer Science*, pages 433–442. Springer Berlin Heidelberg.

Huiskes, M. J. and Lew, M. S. (2008). The MIR Flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA. ACM.

Huiskes, M. J., Thomee, B., and Lew, M. S. (2010). New trends and ideas in visual concept detection: The MIR Flickr retrieval evaluation initiative. In *MIR '10: Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval*, pages 527–536, New York, NY, USA. ACM.

Jaimes, A., Chang, S.-F., and Loui, A. C. (2003). Detection of non-identical duplicate consumer photographs. In *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, volume 1, pages 16–20. IEEE.

Jegou, H., Douze, M., and Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *Computer Vision–ECCV 2008*, pages 304–317. Springer.

Jinda-Apiraksa, A., Vonikakis, V., and Winkler, S. (2013). California-nd: An annotated dataset for near-duplicate detection in personal photo collections. In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, pages 142–147. IEEE.

Kim, H.-S., Chang, H.-W., Lee, J., and Lee, D. (2010). BASIL: effective near-duplicate image detection using gene sequence alignment. In *Advances in Information Retrieval*, pages 229–240. Springer.

Nister, D. and Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2161–2168. IEEE.

Niu, X.-m. and Jiao, Y.-h. (2008). An overview of perceptual hashing. *Acta Electronica Sinica*, 36(7):1405–1411.

Over, P. (2014). TREC Video Retrieval Evaluation: TRECVID. http://trecvid.nist.gov/.

Pollock, K. H., Nichols, J. D., Brownie, C., and Hines, J. E. (1990). Statistical inference for capture-recapture experiments. *Wildlife monographs*, pages 3–97.

Vonikakis, V., Jinda-Apiraksa, A., and Winkler, S. (2014). Photocluster - a multi-clustering technique for near-duplicate detection in personal photo collections. In *Proc. of the 9th International Conference on Computer Vision Theory and Applications*, pages 153–161.

Won, C. S., Park, D. K., and Park, S.-J. (2002). Efficient use of mpeg-7 edge histogram descriptor. *Etri Journal*, 24(1):23–30.

Zezula, P., Amato, G., Dohnal, V., and Batko, M. (2006). *Similarity search: the metric space approach*, volume 32. Springer.