Quantifying the Specificity of Near-Duplicate Image Classification Functions

Richard Connor[1] and Franco Alberto Cardillo[2]

[1] *Department of Computer and Information Sciences,*
University of Strathclyde, Glasgow, G1 1XH, Scotland
[2]*Istituto di Scienza e Tecnologie dell'Informazione,*
Consiglio Nazionale delle Ricerche, Pisa, Italy
richard.connor@strath.ac.uk, franco.alberto.cardillo@isti.cnr.it

Abstract: There are many published methods for detecting similar and near-duplicate images. Here, we consider their use in the context of unsupervised near-duplicate detection, where the task is to find a (relatively small) near-duplicate intersection of two large candidate sets. Such scenarios are of particular importance in forensic near-duplicate detection. The essential properties of a such a function are: performance, sensitivity, and specificity. We show that, as collection sizes increase, then specificity becomes the most important of these, as without very high specificity huge numbers of false positive matches will be identified. This makes even very fast, highly sensitive methods completely useless. Until now, to our knowledge, no attempt has been made to measure the specificity of near-duplicate finders, or even to compare them with each other. Recently, a benchmark set of near-duplicate images has been established which allows such assessment by giving a near-duplicate ground truth over a large general image collection. Using this we establish a methodology for calculating specificity. A number of the most likely candidate functions are compared with each other and accurate measurement of sensitivity vs. specificity are given. We believe these are the first such figures be to calculated for any such function.

1   INTRODUCTION

In forensic image detection, it is commonly required to determine if one large image collection contains images for which near-duplicates exist within another large collection. For example, the UK National Crime Agency has a collection of approximately 10^7 known child abuse images. A media device seized from a suspect may contain 10^6 images, but perhaps only 10^3 of these would be of child abuse and thus a subject for prosecution. A knowledgeable suspect will have performed minor visual editing of these images, making them detectable only by near-duplicate finding functions.

Finding images similar to one other, from within a huge collection, is nowadays a (relatively) solved problem, and has been shown to scale to up to collections of well over 10^{10} images. However almost all such results are set in a human-guided search scenario, where a single image is presented to the system and a human is available to pick the best results from those returned. Here we examine quite a different scenario, where two moderately large collections are the input, and the desired output is the set of images of which near-duplicates exist across the two collections. This causes new problems, due to the inherent squaring: if both collections contain a relatively modest 10^6 images, then there are 10^{12} pairs to consider, rather more comparisons than required to compare a single image against Google's indexed image collection.

The performance issues can be handled by various means, including using the obvious parallelisation inherent in the problem. In this article we concentrate on a more subtle problem, that of the precision, or specificity, of the classification function. As there are a really huge number of pairs of images to consider, the specificity must be incredibly high to avoid very large numbers of false positives. For example, specificity of $1 - \frac{1}{10^6}$ would, in most contexts, give excellent precision; but set in this example context would give 10^6 false positives, impossible for any human to check. This is greatly in contrast with human-guided search, where specificity as low as 0.5 is quite acceptable.

Finding near-duplicate images in this context requires a *classification* function; that is, a boolean similarity function *s* which allows the detection of the near-duplicate intersection: from collections X and Y, the near-duplicate intersection is defined as the set of pairs $\{x, y \leftarrow X, Y, s(x,y)\}$. As similarity functions are generally numeric, a decision must be made for a threshold at which to apply the function, giving acceptable tradeoffs for sensitivity and specificity.

Specificity however is very difficult to measure, requiring as it does a very large collection with a known ground truth. Such a collection has been established for images in one context, in (Connor et al., 2015; Connor, 2015), which gives a quantified estimate of the near-duplicate ground truth for a collection of one million images. Using this collection, we show a methodology for quantifying the specificity of near-duplicate finders, and give results for some of the most likely candidate functions.

2 BACKGROUND

Finding a pair of images, one of which has been created by applying minor transformations to the other, is not an easy problem. Such transformations, available in off-the-shelf image processing software, include changes to contrast, brightness, colour, texture, and sharpness. It is easy for the human brain to see that two such images are obviously the same modulo such changes, but these are exactly the properties that most similarity functions rely on for a mechanised quantification of similarity.

Implicitly, the context of testing for similarity function performance is usually a human user searching a very large collection for images that are most similar to a given reference image. As such, efforts tend to be concentrated on the performance and recall of nearest-neighbour search, and success is generally measured in terms of how many correct results can be obtained from a huge collection within a short time.

Our context of interest is quite different however. Increasing pressure is being put on Internet Service Providers, social network and search engine providers to filter image and video content that is being used illegally, for example the portrayal of child abuse or content which is subject to copyright. Such content is always subject to minor changes, for a number of reasons but increasingly including deliberate attempts to evade detection. Therefore near-duplicate rather than duplicate detection is increasingly required.

In these and other contexts, the requirement is to automatically search large numbers of images against a reference collection which is also large, for example 10^6 to 10^7 images or keyframes. When a possible near-duplicate is detected, the two images then require to be tested by human inspection. Efficiency and high-quality semantic matching are requirements, but false positive detection becomes a much bigger issue. As the number of comparisons required is huge, and the majority of candidates have no match, false positive detection must be an incredibly rare event.

Figures 1 and 2 show some perhaps surprising motivating examples where very small distances have been observed over very different images, as a result of a very low false positive probability being applied over a very large sample population. In both of these cases, the image pairs shown have very small distances measured according to a particular similarity function: that is, distances that were smaller than the majority of true near-duplicate pairs. These examples were found in a collection of 10^{12} image pairs: such coincidental matches are quite likely to occur even if their probability is as low as 10^{-12}.

The contribution of this paper is to show a way of performing specificity measurement useful for this context, and to give some early analysis of various near-duplicate finding mechanisms.

Figure 1: Very similar images, as determined by edge histograms

Figure 2: Very similar images, as determined by colour histograms

3   RELATED WORK

We are not aware of any other work which attempts to measure comparative sensitivity and specificity for different image similarity functions. We believe one reason for this is the lack of any large collection of images with a known ground truth of similarity, a lack mentioned by various authors, for example: *"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), and the same lack is noted in (Jinda-Apiraksa et al., 2013). (Vonikakis et al., 2014) note *" Although the target application of this dataset is image retrieval, it was selected due to the lack of other appropriate datasets [. . .]"*.

There is also only a little work which objectively compares different image similarity functions. Royo (Ventura Royo, 2010) gives a comparison of different MPEG-7 techniques for image search, however, in common with most published work in this domain, the notion of success is based on correct retrieval from a relatively small collection of known images. Interestingly, the author finds that MPEG-7 Colour Structure gives the best performance in this context, while we find it by far the worst of those tested. The main point is that the issues addressed here are defined by the scale of the search, and comparative studies over small collections of images do not give useful results.

Douze et al. (Won et al., 2002) compare the GIST (Oliva and Torralba, 2001) image characterisation with a "bag-of-features" approach, and find it superior for near-duplicate images among other things. They test over significant sizes of image collection by adding their ground truth of 'similar' images to large 'distractor' sets, including MIR-Flickr. However they generate near-duplicate images from a relatively small, chosen set, through programmed quality loss and cropping, and also apply strong transformations which are not covered by our definition of near-duplicate. The use of generated near-duplicate images weakens the results, as for any known transformation it is relatively straightforward to predict which similarity functions will perform well.

Foo et al. (Foo et al., 2006) gave one of the earliest treatments of the issue of near-duplicate image finding as a subject in its own right, and defined two categories of near-duplicate images: *IND* and *NIND*. *IND* images are "derived from the same digital source after applying some transformations", and *NIND* images "share the same scenes and objects". Here we concentrate primarily on the *IND* category which is most appropriate for our defined problem domain.

Our measurement of image similarity functions

for this purpose is based over the MIR-Flickr collection of one million images (Huiskes and Lew, 2008; Huiskes et al., 2010). This dataset 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. 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 of near duplicate detection algorithms on large image datasets.

Using this collection, the Mir-Flickr Near Duplicate (MFND) classification (Connor et al., 2015; Connor, 2015) identifies three sets of clusters which occur within the original set of one million images: duplicate clusters, *IND* near-duplicate clusters, and *NIND* near-duplicate clusters. The *IND* collection comprises 1,958 clusters containing a total of 4,071 images, the majority of clusters containing only two elements. Using population statistics, the authors have shown that these are almost all of the *IND* pairs that exist within the whole set. Based on three relatively independent near-duplicate finding functions, an estimate is made of the total number which exist, including those not found. Furthermore, the standard error of this estimate has also been established at under 0.02. Therefore if an image is chosen randomly from outside these three clusters, there is a very low probability that a near-duplicate will exist within the rest of the collection. We have used this property in order to establish the specificity measurements given in this article.

There are many image similarity functions described in the literature. In general, these can be classified into those relying on global, and local, features. Here we report only the class of global feature functions, as we believe that these are most likely to be suitable for our specific problem domain. This is a questionable assumption that we are investigating further, although backed up by results in (Won et al., 2002) and (Chum et al., 2007). Table 1 summarises the different characterisations studied.

In many cases a particular distance metric is specified as a part of the published mechanism. For example (Won et al., 2002) prescribes the use of L_1 (Manhattan) distance, and most assume that L_2 (Euclidean) distance is the best measure of distance between characterisations. However we have found these presumptions are most often wrong, certainly in this context, and we have tested all characterisations over a number of different metrics.

In some cases, optimisation mechanisms are bun-

Table 1: Image characterisations used

Abbreviation	Characterisation	Reference
Cs	MPEG-7 Colour Structures	(Bober, 2001)
Csl	MPEG-7 Colour Structure Layout	(ISO-15938,)
Eh	MPEG-7 Edge Histograms	(Won et al., 2002)
Ghch	Global Hierarchical Colour Histograms	(Chum et al., 2007)
Gist	GIST	(Oliva and Torralba, 2001)
Ht	MPEG-7 Heterogeneous Textures	(Bober, 2001)
Phash	Perceptual Hashing	(Niu and Jiao, 2008)

dled with the extraction description, for example (Niu and Jiao, 2008) extracts a bitmap for comparison with Hamming distance, and (Chum et al., 2007) uses locality-sensitive hashing techniques over the extracted representations. Rather than do this we maintain the original representations and apply general metrics over them; optimisation techniques are best treated separately.

Finally, three of these characterisations rely upon the image colour palette; *Cs* very strongly, and *Csl* and *Ghch* much less so, as despite their names the most significant information in these is extracted according to pixel intensity rather than colour. However, results for *Cs* are very bad, and based on this observation we repeated tests on *Csl* and *Ghch* using only intensity, rather than colours. As they were in all cases better, we report only these variants here. From observation, many *IND* near-duplicate images in our reference set have been produced by making changes to the colour palette, which explains this observation.

4 DEFINITIONS AND CONTEXT

We assume that any near-duplicate finder is based on a positive numeric function D over any two images. Normally D will be a proper distance metric to allow scaling of the search, but this is not an essential semantic property.

To run an unsupervised search, it is necessary to use D as a classification function. To achieve this, a distance threshold t must be chosen to be used in conjunction with D to form a predicate function D_t over image pairs, such that $D_t(x,y) = D(x,y) \leq t$.

The problem domain can then be characterised as the requirement to find the near-duplicate intersection of two image sets X and Y, based on a concepual near-duplicate relation ND, where this is defined as the set of pairs $X \cap_{ND} Y$ where $(x,y) \in X \cap_{ND} Y \iff ND(x,y)$.

Normal definitions of sensitivity and specificity for a threshold function D_t can now be defined as conditional probabilities:

$$sens_{D_t} = \mathcal{P}(D_t(x,y)), \text{given } (x,y) \in X \cap_{ND} Y$$

$$spec_{D_t} = \mathcal{P}((x,y) \notin X \cap_{ND} Y), \text{given } \neg D_t(x,y)$$

In general, as the search threshold t is increased, the sensitivity increases and the specificity decreases.

As the collections become larger, the specificity becomes increasingly important. After execution of the unsupervised process, the number of true positive matches found will be

$$sens_{D_t} \cdot |X \cap_{ND} Y|$$

and the number of false positives will be

$$(1 - spec_{D_t}) \cdot |X| \cdot |Y|$$

To put this into a realistic context, a typical situation for the detection of child abuse images from a seized hard drive is

$$|X| = 10^5, |Y| = 10^7, |X \cap_{ND} Y| = 10^3$$

so apparently excellent figures of, for example,

$$sens_{D_t} = 0.999, spec_{D_t} = 0.999$$

leads to the detection of almost all the 10^3 true positives, but these will be impossible to find among 10^9 false positives. In fact, in this scenario, the specificity requires to be as high as $1 - 10^{-9}$ before only around half of the detected pairs will be true positives.

To investigate such probabilities requires analysis over large image sets with known ground truths, where the combination of very rare events and very large populations can be quantified.

5 METHODOLOGY

The MFND benchmark set defines three sets of image clusters, which are subsets of the one million images in the MIR-Flickr set: one of identical, one of IND and one of NIND images. Results here use the IND set as this is more likely to be complete, however

we have run the tests reported here using both IND and NIND sets and the outcomes are quantitatively indistinguishable.

To measure the sensitivity of a given metric across a range of thresholds, a set of pairs comprising the first two images from each cluster was used, giving a set of 1,958 pairs of IND images. A histogram of the distances is constructed, and a cumulative probability density function constructed from this to give sensitivity for each D_t across the range of different t values for each similarity function D.

To measure specificity of a given metric, a set of 5,000 images was randomly selected, ensuring that none of these were in any of the IND, NIND, or duplicate clusters defined by the benchmark collection. It is thus a safe assumption that none of these images have a near-duplicate match within the collection. For each of these images, its *nearest neighbour* from within the 1M collection, along with the distance to this, was determined. Note that in principle this requires 5×10^9 distance calculations, and metric indexing techniques were used to achieve tractability.

As nearest neighbours are calculated, the smallest of the nearest-neighbour distances is therefore the smallest of approximately 5×10^9 distances measured across the two sets. As this distance relates two images which are not near-duplicates, this gives the specificity of the function D_t, where t is this smallest distance, as $1 - (2 \times 10^{-10})$ for these two sets of images.

In measuring specificity we make the simplifying assumption that, at least for the smallest few nearest neighbours measured from the set, these distances are the smallest from the whole set of potential distances; that is, not allowing for a single image to have a second neighbour which is a smaller absolute distance than another's nearest neighbour. This is likely to be the case, but more importantly it also captures a more useful figure for use in unsupervised detection scenarios, where a threshold-limited nearest neighbour search will be conducted.

We calculate results in terms of a cumulative probability density function, therefore a threshold which admits the five smallest nearest-neighbour distances, i.e. a cumulative density of 0.001 over the 5,000 test distances, corresponds to a specificity of $1 - 10^{-9}$. The *cdf* value of 0.01 corresponds to a specificity of $1 - 10^{-8}$ which as explained earlier is around the smallest useful threshold for our problem context: when comparing 10^5 images against 10^7, around 10^4 false positives would result, this number varying with the product of the two collection sizes.

To give an example of this analysis, Figure 3 shows simple histograms giving the outcome of using
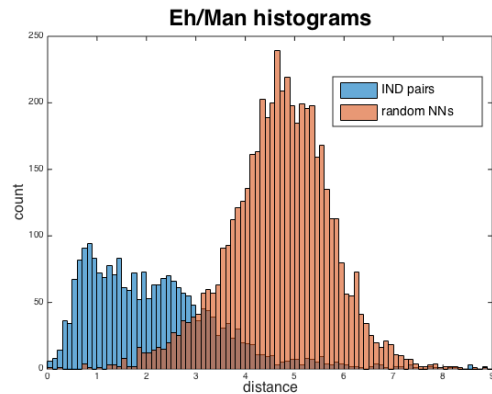

Figure 3: Eh/Man Histograms

Manhattan distance over MPEG-7 edge histograms (Won et al., 2002). The histograms show the distribution of distances over (1) known *IND* near-duplicate image pairs, and (2) 5,000 randomly selected images and their nearest neighbours from within the set of 1M images – this latter of course is shifted a long way to the left of a histogram showing the distribution of distances of randomly selected image pairs.

As can be seen, there is significant overlap between the histograms. Figure 4 shows the same data displayed as the deduced cumulative probability density functions. From these, it can be seen that using this function with a threshold of around 2 will give a specificity of around $1 - (10^{-8})$ and a sensitivity of around 0.5, but to achieve an order of magnitude improvement in specificity would allow a sensitivity of only around 0.1.

In order to allow comparisons of the different characterisation and metrics tested, these results are shown plotted in ROC curves, to show the essential tradeoff between sensitivity and specificity as the search threshold is increased. Figure 5 shows the same data again plotted as ROC graphs, which is the form we will use from now on to present results of the different functions tested.

6 MEASUREMENTS

For each different characterisation shown in Table 1, we have applied the following proper distance metrics: Manhattan distance, Euclidean distance, Cosine distance[1] and, where possible, Structural Entropic Distance (*SED*). There are of course many other potential metrics which could be tested.

---

[1]the angle between the vectors rather than the complement of its cosine, which is not a proper metric
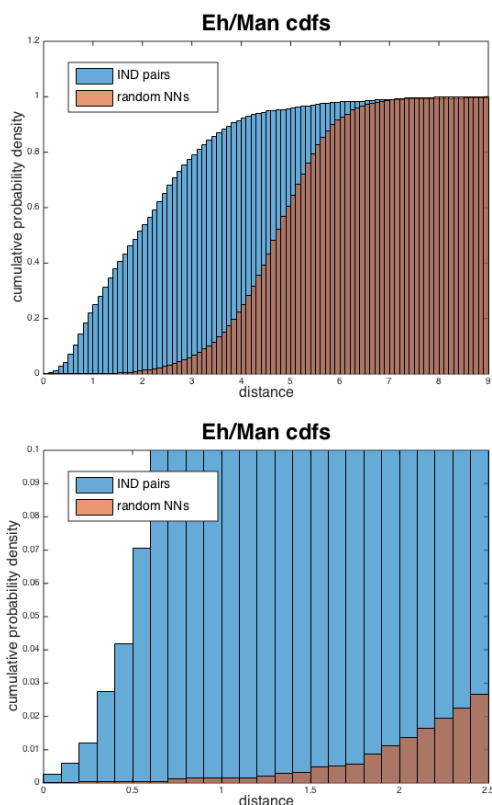
Figure 4: Eh/Man Cumulative Probability Density (the lower graph is a magnification at the origin)

Figure 5: Eh/Man ROC graphs. The lower graph is the extreme left-hand part of the upper, as required for this context.

SED is the distance metric defined in (Connor et al., 2011) and refined in (Connor and Moss, 2012) for use in general vector spaces. It is defined over probability distributions, and as such can only be applied to characterisations all of whose output values are positive - notably not including those defined by discrete cosine transform, namely *Csl* and *Phash*. For all characterisations comprising only positive numeric values, they are normalised to sum to 1 for this purpose. *SED* gives the same ranking of outcomes as the better-known Jensen-Shannon distance (Lin, 1991), but is used here as it has better efficiency properties for querying metric spaces in this context (Connor and Moss, 2012); the qualitative results shown here would be exactly the same for both metrics.

Finally it is important to stress that we have examined only the semantic properties of each image similarity function, rather than their relative search efficiency which varies very widely, as does their extraction time. As mentioned earlier, there are many ways of optimising both of these and an objective comparison of performance is not useful in simple terms.
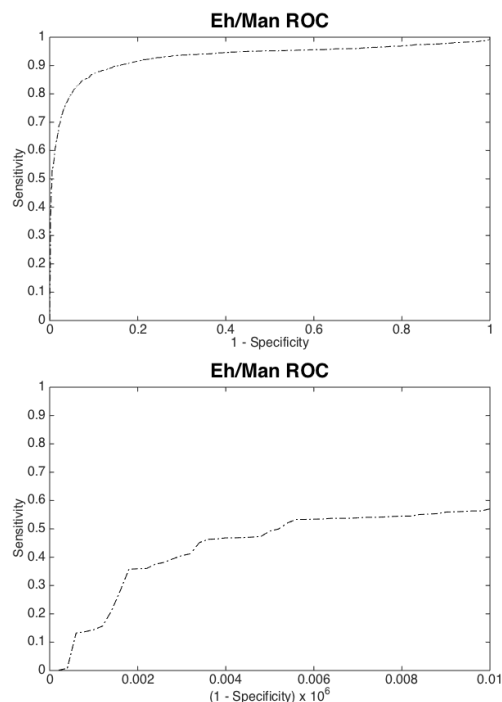
7 RESULTS

Figures 6 to 11 show results over the same axes for six of the seven characterisations tested. In each case the specificity range is restricted from 1 to $1 - 10^{-8}$ as previously explained. As can be extrapolated from these graphs, at this point the sensitivity improvement given by a decrease in specificity becomes marginal.

The missing characterisation, MPEG-7 Colour Structures, performs so badly it is not shown, achieving a sensitivity in this range of less than 0.02 with any metric. The importance of measuring these over very large collections is emphasised by the finding by other authors that this is the best of the MPEG-7 characterisations over small test collections.

The result of these tests is clear from inspecting the graphs: of the 22 methods tested for near-duplicate classification, the best compromise between sensitivity and specificity is the use of Structural Entropic Distance/Jensen-Shannon Distance over the GIST image representations, which gives a sensitivity of around 0.8 for a specificity as high as $1 - 10^{-9}$.

Applying this in the context of a scenario with two collections each of 10^6 images, with a near-duplicate intersection of 10^3 images, then querying at the appropriate threshold will return 800 of the near-duplicate image pairs, and 1,000 false positive pairs.
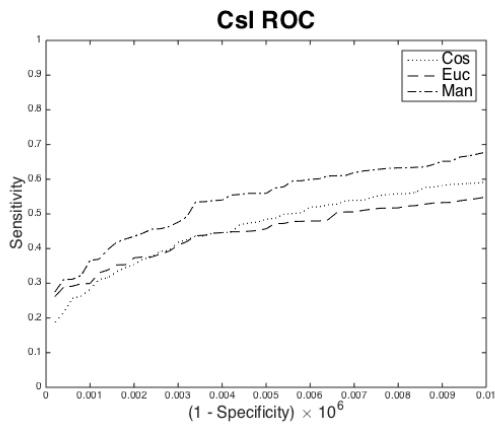
Csl ROC
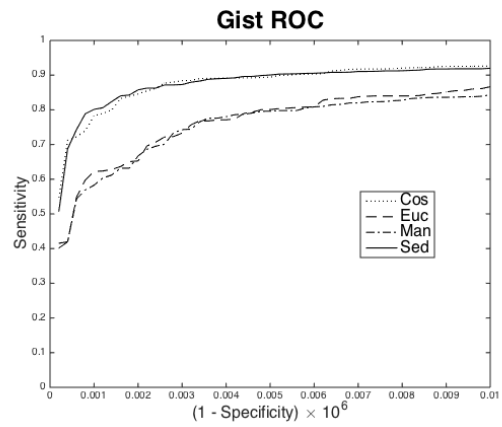

Figure 6: Csl ROC graph for three metrics

Eh ROC

Figure 7: Eh ROC graph for four metrics

Ghch ROC

Figure 8: Ghch ROC graph for four metrics

Gist ROC

Figure 9: Gist ROC graph for four metrics

Ht ROC
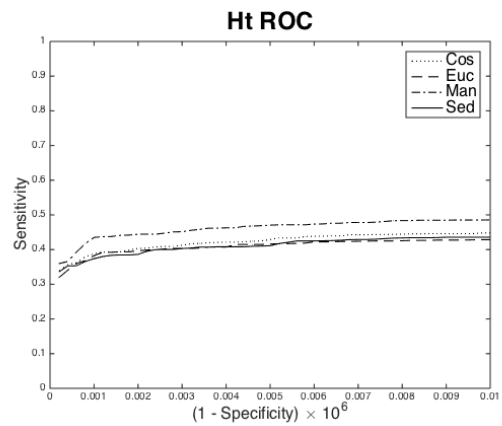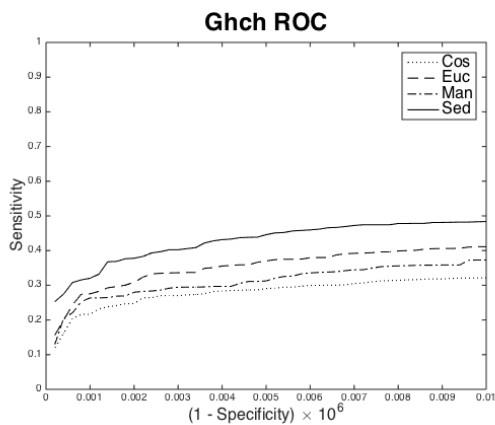

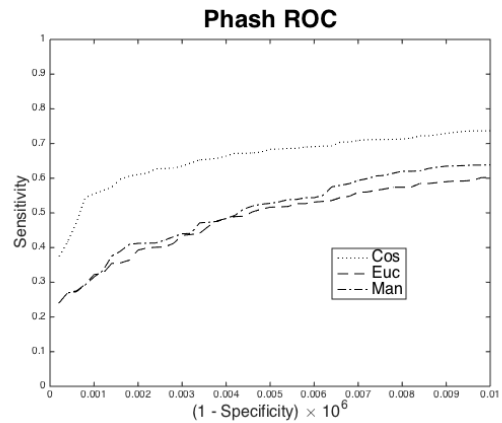Figure 10: Ht ROC graph for four metrics

Phash ROC

Figure 11: Phash ROC graph for three metrics

By comparison, for example, using perceptual hashing with Euclidean distance will return 5,000 false positive pairs for every 500 true pairs.

8 CONCLUSIONS

The main value of this work is to provide a re-usable methodology by which different near-duplicate finders can be compared for specificity. The particular value of our method is that it is based upon a

large collection of "naturally" selected images which happen to contain near-duplicates, rather than a constructed set. The resulting ground truth should not therefore be biassed towards any particular function.

Validation of the results however is a further challenge, at least until such time as another large image set with a near-duplicate ground truth is identified. We are currently working with the *CoPHiR* collection (Bolettieri et al., 2009) (10^8 images) to establish whether the figures produced here are consistent.

The variation among the different distance metrics is a novel observation. Characterisations are normally used with either L_1 or L_2 distance, whereas in the majority of cases either Cosine or *SED/JSD* performs best. These metrics give a closer match according to the correlation of values within the characterisations, rather than differences in their absolute magnitude. However the differences among all the characterisations do not seem to suggest any general rules about the best metric to use in different contexts, which requires further investigation.

ACKNOWLEDGEMENTS

REFERENCES

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

Bolettieri, P., Esuli, A., Falchi, F., Lucchese, C., Perego, R., Piccioli, T., and Rabitti, F. (2009). Cophir: a test collection for content-based image retrieval. *CoRR*, abs/0905.4627.

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., Cardillo, F., MacKenzie-Leigh, S., and Moss, R. (2015). Identification of mir-flickr near-duplicate images. In *10th International Conference on Computer Vision Theory and Applications*.

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.

ISO-15938. Mpeg-7 multimedia content description interface.

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.

Lin, J. (1991). Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145–151.

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

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175.

Ventura Royo, C. (2010). Image-based query by example using mpeg-7 visual descriptors.

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.