Perceptual hashing and its future in disinformation detection

Edinburgh Napier
UNIVERSITY

By Clara O’Callaghan

01 Introduction

Perceptual image hashing is the process of encoding an image based on visual features a human would find signficant and output a hash based on those features. By comparing hashes of two images, you can tell how different one image is from another. This type of hashing is used commercially, for example in Reverse Image Search. This experiment aimed to identify if minor changes such as small edits used to disinform the audience would create a large enough change in the hash. Additionally, would these changes act logically, and the hash difference be directly dependent on the size of the change?

02 Methodology

This experiment used 44 thousand images across 2 subsets of the defacto[1] dataset (copy-move and inpainting). It focused on localised (small) changes across the images.

Methodology ->

1. Define the algorithms used (pHash, dHash, wHash)
2. Calculate the percentage of the image that was edited, the location, the hashes and the respective hamming distances (HDs)
3. Identify images with a low percentage of edit and high HD or high percentage change and low HD
4. Remove any outlying data [2][3][4]
5. Analyse these images and their prevalence

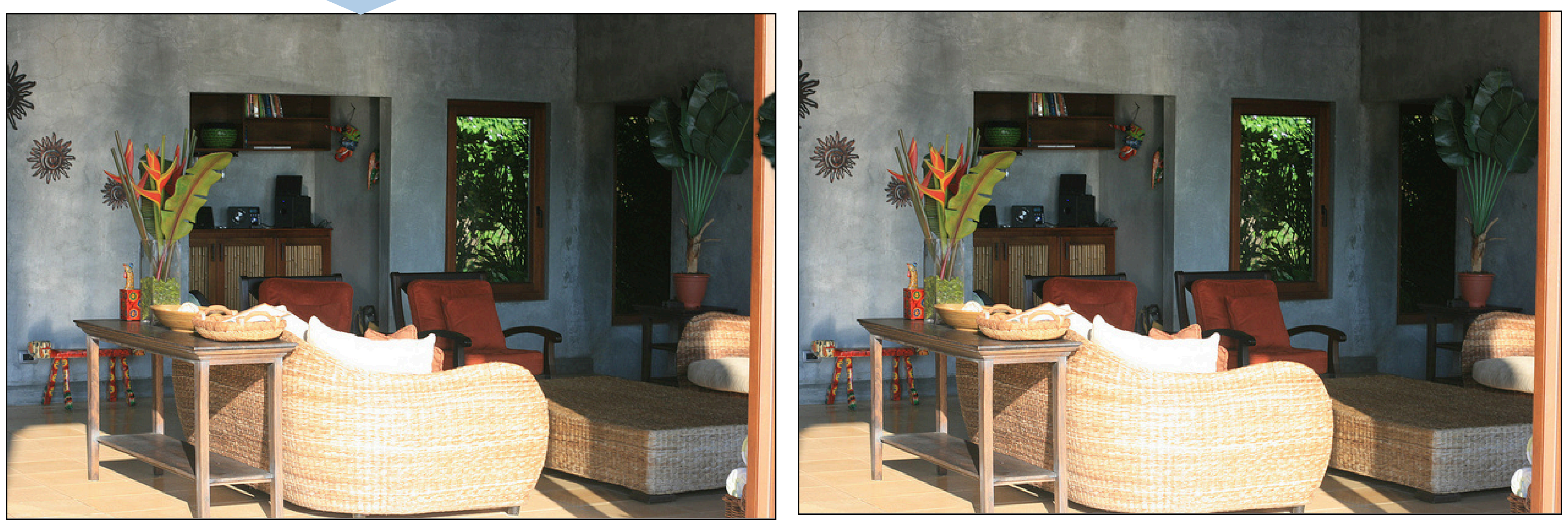
Hamming Distance (HD) the difference bit by bit between two hashes.

Inpainting removing an object in the image and filling the space.

03 Results and Findings

- Some hashing algorithms are more likely to see smaller changes.
- This comes at the cost of sometimes identifying a small change as a major difference.
- For pHash, this seems to happen due to lighting and/or lack of distinct features. This isn’t always obvious in a picture full of features. This image has a (pHash) normalised HD of 0.41.
- For wHash, when an image is low quality, the object is a similar colour to its surroundings, or the structure change of the image is minimal, the removal of an object doesn’t affect the hash much[3].

Can you spot the difference?



- There were 14,052 images wHash didn’t notice a change in that the others did. Of these 262 images had an edit of over 5%!
- Below 5%, the type of transformation does not make much of a difference. Overall, copy-moving an object is more likely to cause a larger distance.
- Between 5% and 35% of the image being edited change the difference between the two increase.
- At higher changes, inpainting is identified at a bigger impact than copymove.
- The size is a pretty good indicator of difference for dHash and wHash in inpainting, less so in copy move.

Disinformation fake news that is created or shared knowing it isn’t real.

04 Key Takeaways

1. Copy move edits are likely easier to distinguish due to how the features are extracted. The object duplication makes it easier to identify as it retains many of its original characteristics.
2. The prevalence of edited images and their purpose is difficult to determine. There are 14 billion images uploaded daily [5]. This isn’t plausible to analyse (even with Machine Learning).
3. To confidently judge the accuracy requires human-led checking. However, people are very bad at knowing if an image is edited [6].
4. The flaws of any individual algorithm are likely to make the false negative/positive rates higher than is plausible for any system.
5. The difference between an edit to manipulate and an set of content-preserving changes is very small [7][8].
6. Making a small edit in an almost unnoticeable way, or a larger edit in the right place that bypasses the checks would be easy to carry out.

05 Conclusion

Overall, some application of perceptual image hashing can be used to identify small changes within an image, and thus, some disinformation campaigns utilising relatively simple edits that change the context of the image can go unnoticed under the threshold for minor preserving edits like compression or colour saturation. Conversely, one small change in a selective place can considerably change the hash. Some hash algorithms perform better than others, but the technology to progressively identify edits and not identify content-preserving changes in the crossfire is not robust enough for commercial use.

Future work could look at combining algorithms as well as identifying the most common changes within current disinformation to further inform research on the indicators of edits for disinformation and just for aesthetics. There is also a need for education in identifying manipulations. Future human-based research could look at developing a framework for educating users on the markers of editing.

[1] G. Mahfoudi, B. Tajini, F. Retraint, F. Morain-Nicolier, J. Luc, and M. Pic, “DEFACTO : Image and Face Manipulation Dataset,” 2019. [Online]. Available: https://defactodataset.github.io [2] E. Klinger and D. Starkweather, “pHash ,” https://phash.org/. [3] “Overview of Perceptual Hashing Technology Contents,” Nov. 2021. [4] A. Ramos, “Introduction to Perceptual Hashes: Measuring Similarity,” https://apiumhub.com/tech-blog-barcelona/introduction-perceptual-hashes-measuring-similarity/. [5] Matic Broz, “How many pictures are there in 2024?,” https://photutorial.com/photos-statistics/. [6] S. J. Nightingale, K. A. Wade, and D. G. Watson, “Can people identify original and manipulated photos of real-world scenes?,” Cogn Res Princ Implic, vol. 2, no. 1, Dec. 2017, doi: 10.1186/s41235-017-0067-2. [7] P. Samanta and S. Jain, “Analysis of Perceptual Hashing Algorithms in Image Manipulation Detection,” in Procedia Computer Science, Elsevier B.V., 2021, pp. 203–212. doi: 10.1016/j.procs.2021.05.021. [8] A. Hadmi, W. Puech, B. Ait, E. Said, and A. A. Ouahman, “Perceptual Image Hashing,” 2012. [Online]. Available: www.intechopen.com