International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:12, No:5, 2018

Quick Similarity Measurement of Binary Images via Probabilistic Pixel Mapping

Adnan A. Y. Mustafa

***Abstract*—**In this paper we present a quick technique to measure the similarity between binary images. The technique is based on a probabilistic mapping approach and is fast because only a minute percentage of the image pixels need to be compared to measure the similarity, and not the whole image. We exploit the power of the Probabilistic Matching Model for Binary Images (PMMBI) to arrive at an estimate of the similarity. We show that the estimate is a good approximation of the actual value, and the quality of the estimate can be improved further with increased image mappings. Furthermore, the technique is image size invariant; the similarity between big images can be measured as fast as that for small images. Examples of trials conducted on real images are presented.

***Keywords*—**Big images, binary images, similarity, matching.

I. Introduction

IMAGE matching is a fundamental task in robot and computer vision. It is at the core of many problems such as template matching, image registration, motion detection and tracking, image retrieval, as well as many other problems. At the core of any matching process is the similarity criteria employed to measure image closeness. In the case of image matching, the similarity measure is maximized to find the best image match when matching a query image to a database of candidate images. Alternatively, a dissimilarity measure can be employed.

Many similarity measures have been proposed and developed, such as correlation based methods or mutual information based methods; however they all suffer from one major handicap; they all require that the whole image be processed to measure the similarity between two images. This is a serious problem which has not been properly addressed, as it implies that processing time will continue to increase as image size increases. With standard image sizes doubling roughly every 7 years and higher image resolution in greater demand in many applications (particularly in medical applications), the matching process becomes an extremely time consuming and expensive process.

In this paper we present a technique to measure image similarity between images quickly. The technique is image size invariant, giving it an important advantage over other methods. The technique uses a probabilistic approach and is based on probabilistic matching models for binary image matching: the Probabilistic Matching Model (PMM) [1] and the more general PMMBI [2]. By employing these models and recording how fast the dissimilarity between images can be detected, the amount of similarity between images be quickly estimated to a good degree.

Adnan A. Y. Mustafa is with the Department of Mechanical Engineering at Kuwait University, Kuwait (phone: 965-2498-7117; fax: 965-2483-2417; e-mail: adnan.mustafa@ku.edu.kw).

II. Literature Review

The literature is rich with numerous measures that can be employed to measure the similarity between binary images; the Jaccard distance [3], the Sokal-Michener Coefficient (commonly known as the Simple Matching Coefficient) [4], the Hamming distance [5], the Cosine similarity [6], Pearson's correlation coefficient [7], image subtraction methods (e.g. the sum of the absolute difference method (SADM)) [8], Mutual Information (MI) distances [9], and the Gamma binary distance (γ) [10]. Other binary similarity measures and distances have also been cited in the literature [11]. It is important to note that all of these methods require the whole image be processed in order to measure the similarity between two images.

III. Related Work

In this section we summarize previous related work necessary for the understanding of the work presented. Image closeness, the Gamma binary distance and the PMMBI model are reviewed.

A. Binary Image Closeness

Binary images are classified as similar or dissimilar [12]. Furthermore, two similar images are classified as either: 1) exact-similar, if all pixels mapped from one image to the other image have the same intensity values; or 2) inverse-similar, if they have the inverted intensity values at each pixel. If the images are not similar, then they are dissimilar. Furthermore, 1) if the dissimilarity between two images is maximized then the images are classified as distinct-dissimilar, 2) if the dissimilarity is not maximized then the images are classified as quasi-dissimilar. This classification can be easily performed based on the Gamma binary similarity distance (γ), described next.

B. The Gamma Binary Distance

The Gamma binary similarity measure (γ) measures the amount of similarity and concurrence between two binary images [10] [13]. Given two images **u** and **v**, γ is defined as,

$$\gamma(\mathbf{u},\mathbf{v}) = |1 - 2P_o((Z = \mathbf{u} \oplus \mathbf{v}) = z)|,\ z \in \{0,1\} \qquad (1)$$

where $\oplus$ γ is the exclusive-or operation and P_o denotes the probability mass function of the image intensities. As a result,

$\gamma \in [0,1]$,

- A value of $\gamma = 0$ between two images implies that the images are distinct-dissimilar images.
- Values in the range of $0 < \gamma < 1$ imply that the images are classified as quasi-dissimilar images.
- A value of $\gamma = 1$ implies that the images are classified as similar images.

Based on the values of γ, quasi-dissimilar image pairs can be classified into finer detail as detailed in [13]. Note that:

- Image pairs with $\gamma < 0.01$ are considered to be distinct-dissimilar image pairs.
- Image pairs with $0.99 \le \gamma < 0.999$ are called *near-duplicate* images.
- Image pairs with $0.999 \le \gamma < 1$ are called *near-similar* images.

C. The Probabilistic Mapping Model for Binary Images

In [2], a probabilistic model that predicts the probability of detecting dissimilarity between binary images, called the PMMBI was introduced. PMMBI, governed by (2), predicts the probability of detecting dissimilarity between binary images as a function of p, the number of random mappings mapped thus far, and γ the amount of similarity between them,

$$\Pr(\gamma, p) = 1 - \left(\frac{1}{2}(1+\gamma)\right)^p \left(1 + \left(\frac{1-\gamma}{1+\gamma}\right)^p\right) \quad (2)$$

As earlier stated, γ is a continuous random variable with $0 \le \gamma \le 1$. p is a discrete random variable, with $p \ge 2$. Values of Pr are in the range of $0 \le \Pr \le 1$, where a value of $\Pr = 1$ indicates a 100% probability of detecting dissimilarity, while a value of $\Pr = 0$ indicates the impossibility of detecting dissimilarity (only occurs if the images are 100% similar, i.e. $\gamma = 1$). Hence, the values of Pr represent the confidence of detecting image dissimilarity and thus Pr is also referred to as the *Detection Confidence* (*DC*). The expected value of p as a function of γ, $E[p(\gamma)]$, which is the mean number of mappings required to detect dissimilarity is given by,

$$E[p(\gamma)] = \frac{4}{1-\gamma^2} - 1 \qquad 0 \le \gamma \le 1 \quad (3)$$

Its variance is given by,

$$V[p(\gamma)] = \frac{2(8\cdot\gamma^2 - \gamma^4 + 1)}{(1-\gamma^2)^2} \qquad 0 \le \gamma \le 1 \quad (4)$$

IV. Measuring Similarity

The PMMBI model as stated above predicts the probability of detecting similarity (Pr) or the detection confidence (*DC*) between two binary images, given the number of mappings (p) and the amount of similarity (γ). Fig. 1 shows several curves of $\Pr(\gamma,p)$ versus p for several iso-γ (constant γ) curves, while Fig. 2 shows curves of $\Pr(\gamma,p)$ versus γ for different iso-p curves. It can be seen that for low values of γ, DC approaches unity quickly after only a few mappings (p). For distinct-dissimilar images, DC = 93.8% for 5 mappings, and DC = 99.2% for 8 mappings. At higher values of γ, slightly more mappings are needed, e.g. $DC = 94.4\%$ for 10 mappings and $DC = 99\%$ for 16 mappings. Even for near duplicate images, such as the images shown in Fig. 3, where $\gamma = 0.99$, the number of mappings required for detection are not large; e.g. $DC = 90\%$ for 45 mappings and $DC = 99\%$ for 90 mappings.

A plot of E and the standard deviation $\sigma = \sqrt{V}$ is shown in Fig. 4. From this figure we see that on average less than 8 pixel mappings are required to detect similarity for $\gamma < 0.8$. For larger γ values (i.e. for images that are highly similar) more mappings are required; but nevertheless, the required number of pixel mappings –even though larger than those required for lower γ– constitutes a tiny percentage of the total number of pixels in the images. For example, the near duplicate images of *Leena* shown in Fig. 3 require on average only 200 mappings, regardless of image size!

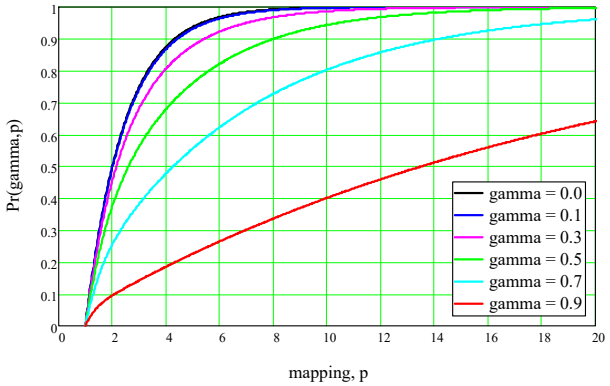

Fig. 1 A plot that displays the variation of $\Pr(\gamma,p)$ as a function of the mapping (p) for several iso-γ curves

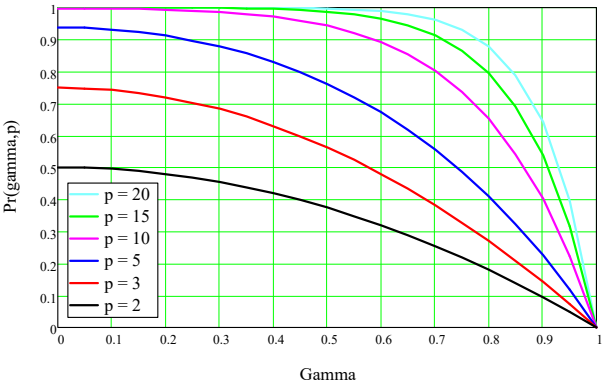

Fig. 2 A plot that displays the variation of $\Pr(\gamma, p)$ versus γ for several iso-p curves

Fig. 3 Highly similar near-duplicate binary images of *Leena* (γ = 0.99). The difference image is also shown. On average 200 mappings are required to detect dissimilarity

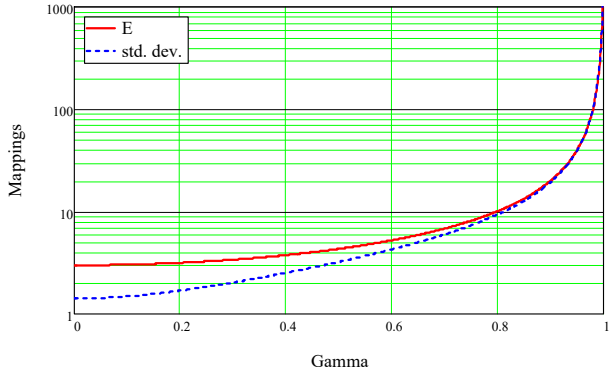

Fig. 4 A plot that displays the variation of $E(\gamma)$ and $V(\gamma)$ vs. γ

From this discussion several conclusions are immediate:

- Detecting dissimilarity between binary images can be performed quickly by randomly selecting a few pixels and comparing their values; there is no need to compare entire images to detect dissimilarity. Even when the images are highly similar, the number of mappings required is extremely small compared to the image size.
- The probability of detecting dissimilarity is not a function of image size; dissimilarity detection quickness (i.e. number of mappings required) is the same whether the image is large or small.
- Given Pr and p, γ can be easily determined. Let *MDN* denote the *Mapping Detection Number*, defined as the number of mappings required to detect a pair of images as being dissimilar. If MDN_μ is the average *MDN* value of n dissimilarity detection trials for a given image pair, then MDN_μ will be close to the expected number of mappings required to detect similarity. In this case, the expected mapping value equation (3) can be used.

For the last point; rearranging (3) and substituting $MDN_\mu = E[p(\gamma)]$ produces an equation of the estimated similarity γ_e as a function of MDN_μ,

$$\gamma_e(MDN_\mu) = \sqrt{1 - \frac{4}{MDN_\mu + 1}} \qquad MDN_\mu > 1 \qquad (5)$$

Thus to measure the similarity between two images we precede as follows:

1. The dissimilarity between the images is detected and the *MDN* is recorded. The detection process is repeated n times.
2. MDN_μ is calculated as the mean of the *MDN* values recorded for the n detection trials.
3. The estimated similarity between the images, γ_e, is then estimated by using (5).

V. Discussion

Fig. 5 shows a plot consisting of five different rounds of (MDN_μ) for an image pair with a similarity of γ = 0.106. The expected number of mappings required for dissimilarity detection by (3) is $E[p(\gamma)]$ = 3.045 mappings, and is shown as the dotted line in the plot. For each round, the MDN_μ value versus the number of trials is plotted. The curves initially fluctuate but then quickly settle down and converge to values close to the expected value of 3.045 mappings. Convergence is so rapid that on average by the 10th trial the error < 10%, by the 28th trial the error < 5%, by the 116th trial the error < 2% and by the 400th trial the error < 1%.

Fig. 6 shows another plot consisting of five different rounds of MDN_μ, but this time for an image pair with a much higher similarity value of γ = 0.782. The expected number of mappings required for dissimilarity detection by (3) is $E(p(\gamma))$ = 9.297 mappings. The curves initially fluctuate but then begin to settle down and converge to values close to the expected value. Convergence is not as rapid as that observed for the previous case because of the higher γ value in this case; on average by the 9th trial the error is less than 10%, by the 247th trial the error settles below 5%, by the 424th trial the |error| < 2%.

It is obvious from the discussion that:

- MDN_μ approaches $E(p(\gamma))$ as predicted by (5) with increasing trials.
- The number of trials required for MDN_μ to approach $E(p(\gamma))$ for higher similarity images is greater than the number of trials required for images with less similarity.

VI. Conclusion

In this paper we have presented a technique to measure image similarity between binary images. The method gives a good estimate of the similarity between binary images by exploiting how quickly images can be detected as being dissimilar by randomly mapping corresponding pixels and detecting dissimilarity; dissimilarity between dissimilar images are detected more rapidly than similar images. With the aid of the PMMBI, which relates the amount of similarity to the number of pixels required to detect dissimilarity, an estimate of the similarity between the images can be obtained. With increasing number of trials, a better estimate of the similarity can be obtained. The method is quick since it only requires the examination of a small portion of the image pixels and not the whole image, as dome by traditional methods. Furthermore, the method is image size invariant and the similarity between big images can be estimated just as quick at that for small images. Examples of trials conducted on real images were presented that show that the estimated similarity

using the theory presented are very close to the actual image similarity values, with errors less than 5%. With increased mappings, the error can be reduced even further to less than 1%. Our future work will concentrate on applying this technique to template matching for big images. We expect that using this approach, the expensive cost of template matching, particularly with large images can be drastically reduced.

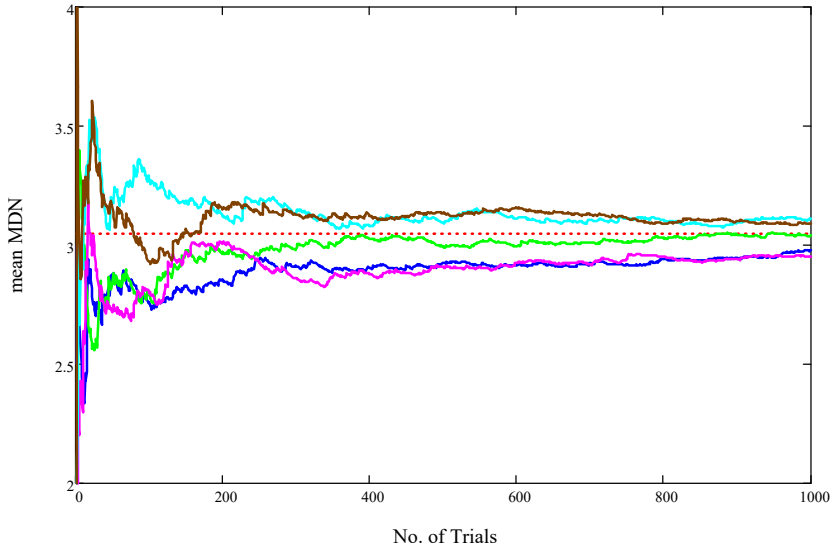

Fig. 5 MDN_{μ} vs. number of trials for five different rounds (image pair with $\gamma = 0.106$). The dotted line represents $E(p(\gamma = 0.106))$

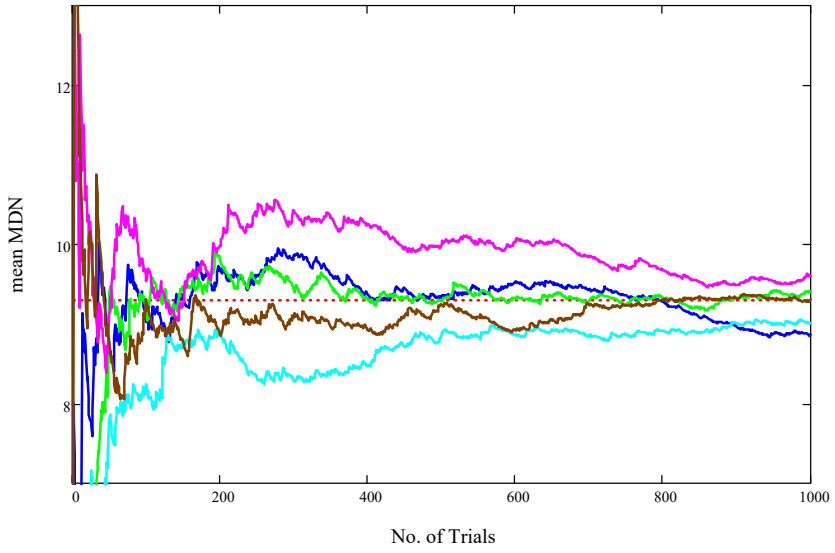

Fig. 6 MDN_{μ} vs. number of trials for five different rounds (image pair with $\gamma = 0.782$). The dotted line represents $E(p(\gamma = 0.782))$

REFERENCES

[1] A. A. Mustafa, "Probabilistic Model for Quick Detection of Dissimilar Binary Images". *Journal of Electronic Imaging*, 24, 5, 2015, pp. 24-53.

[2] A. A. Mustafa, "A Probabilistic Model for Random Binary Image Mapping". *WSEAS Transactions on Systems and Control,* Volume 12, 2017, Art. #34, pp. 317-331.

[3] P. Jaccard, "Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines". *Bulletin de la Société Vaudoise des Sciences Naturelles*, 1901, 37, pp. 241-272.

[4] R. Sokal and C. Michener, "A statistical method for evaluating systematic relationships", Bulletin of the Society of University of Kansas, 1958, 38, pp. 1409-1438.

[5] R. Hamming, "Error detecting and error correcting codes". *Bell System Technical Journal*, 1950, 29, (2), pp. 147–160.

[6] G. Sidorov et al., "Soft similarity and soft cosine measure: Similarity of features in vector space model". *Computación y Sistemas*, 2014, 18, (3), pp. 491-504.

[7] P. Anuta, "Spatial Registration of Multispectral and Multitemporal Digital Imagery Using Fast Fourier Transform Techniques". *IEEE Transactions on Geoscience Electronics*, GE-8, N 4, 1970, pp. 353-368.

[8] D. Barnea and H. Silverman, "A Class of Algorithms for Fast Digital Image Registration". *IEEE Trans. on Computers*, Vol. c-21, N 2, 1972, pp.179-186.

[9] J. Pluim, A. Maintz and M. Viergever, "Mutual-Information-Based Registration of Medical Images: A Survey". *IEEE Transactions on Medical Imaging*, 22, 8, 2003.

[10] A. A. Mustafa, "A Modified Hamming Distance Measure for Quick

Rejection of Dissimilar Binary Images". *International Conference on Computer Vision and Image Analysis*, 2015.

[11] S. Choi, S. Cha, and C. Tappert, 'A Survey of Binary Similarity and Distance Measures'. *Journal of Systems, Cybernetics and Informatics*, 2010, 8, (1), pp. 43-48.

[12] A. A. Mustafa, "Quick Probabilistic Binary Image Matching: Changing the Rules of the Game". *Proc. SPIE* 9971, *Applications of Digital Image Processing XXXIX*, 997112 (September 27, 2016); doi:10.1117/12.2237552.

[13] A. A. Mustafa, "A Probabilistic Binary Similarity Distance for Quick Image Matching". *IET Journal on Image Processing*, submitted for review.