Image similarity index based on moment invariants of approximation level of discrete wavelet transform

P. Premaratne and M. Premaratne

Subjective quality measures based on the human visual system for images do not agree well with well-known metrics such as mean squared error and peak signal-to-noise ratio. Recently, the structural similarity measure (SSIM) has received acclaim owing to its ability to produce results on a par with the human visual system. However, experimental results indicate that noise and blur seriously degrade the performance of the SSIM metric. Furthermore, despite the SSIM's popularity, it does not provide adequate insight into how it handles the 'structural similarity' of images. Proposed is a new structural similarity measure based on the approximation level of a given discrete wavelet decomposition that evaluates moment invariants to capture the structural similarity with superior results over the SSIM.

Introduction: Comparing two images accurately to ascertain whether there is a match or not is essential for many image processing related tasks such as watermarking, compression and content retrieval. Age-old metrics such as mean squared error (MSE) have been used for decades despite its inability to agree with human subjective analysis [1, 2]. Recently, light has been shed on a new metric that seems to agree with the human visual system [2]. The structural similarity measure (SSIM) is supposed to estimate image degradation as a perceived change in structural information. Structural information contains strong inter-dependencies of pixels especially when they are spatially close. These dependencies carry important information about the structure of the objects in the visual scene. The SSIM has been singled out owing to its claim of superiority over the existing metrics [3, 4]. However, it has been observed that the SSIM does not perform well with blurred images [4]. Since a blurred version of an image essentially contains the same structure, the SSIM's inability to measure the structural similarity of blurred images raises an issue as to whether the SSIM does truly look for the structural content. From our research, we have concluded that despite the SSIM's claim of superiority, its ability to compare similar structures is doubtful, as will be demonstrated in this Letter. We have developed a new metric that uses some of the concepts exploited by the SSIM. The new metric demonstrates better performance over the SSIM in blurred images and images corrupted by Gaussian and Salt & Pepper noise.

Structural similarity measure: The SSIM attempts to separate the task of similarity measurement of two images into luminance, contrast and structure [2]. Hence, a similarity measure is defined as:

$$\mathrm{SSIM}(\mathbf{P_1}, \mathbf{P_2}) = l(\mathbf{P_1}, \mathbf{P_2}) \times c(\mathbf{P_1}, \mathbf{P_2}) \times s(\mathbf{P_1}, \mathbf{P_2}) \tag{1}$$

where $\mathbf{P_1}$ and $\mathbf{P_2}$ are the two images being compared and l, c and s stand for luminosity, contrast and similarity measure. Mean μ and standard deviation σ of images $\mathbf{P_1}$ and $\mathbf{P_2}$ are defined as follows:

$$\mu_{P1} = \frac{1}{M \times N} \sum_{y=0}^{M-1} \sum_{x=0}^{N-1} \mathbf{P_1}(x, y)$$

$$\sigma_{P1} = \sqrt{\frac{1}{(M \times N - 1)} \sum_{y=0}^{M-1} \sum_{x=0}^{N-1} (\mathbf{P_1}(x, y) - \mu_{P1})^2}$$

$$\sigma_{P1}\sigma_{P2} = \frac{1}{M \times N - 1} \sum_{y=0}^{M-1} \sum_{x=0}^{N-1} (\mathbf{P_1}(x, y) - \mu_{P1})(\mathbf{P_2}(x, y) - \mu_{P2})$$

Furthermore, l, c and s for images $\mathbf{P_1}$ and $\mathbf{P_2}$ are calculated as follows [3]:

$$l(\mathbf{P_1}, \mathbf{P_2}) = \frac{2\mu_{P1}\mu_{P2} + C_1}{\mu_{P1}^2 + \mu_{P2}^2 C_1}$$

$$c(\mathbf{P_1}, \mathbf{P_2}) = \frac{2\sigma_{P1}\sigma_{P2} + C_2}{\sigma_{P1}^2 + \sigma_{P2}^2 + C_2}, \; s(\mathbf{P_1}, \mathbf{P_2}) = \frac{\sigma_{P1P2} + C_3}{\sigma_{P1}\sigma_{P2} + C_3}$$

Constants C_1, C_2 and C_3 are used for the stability of equations when μ and σ are extremely small. Combining the above definitions, (1) can be expressed as follows when $C_3 = C_2 = 2$ for simplicity [3]:

$$\mathrm{SSIM}(\mathbf{P_1}, \mathbf{P_2}) = \frac{(2\mu_{P1}\mu_{P2} + C_1)(\sigma_{P1P2} + C_2)}{(\mu_{P1}^2 + \mu_{P2}^2 C_1)(\sigma_{P1}^2 + \sigma_{P2}^2 + C_2)}$$

The expression s($\boldsymbol{P}_1$,$\boldsymbol{P}_2$) is a simple function of cross-correlation. It does not contain any notion of structure as structure in an image would represent directionality of objects and how they are organised. Cross-correlation would only capture the similarity of pixels and not structure. This clearly indicates that the SSIM does not evaluate structure and hence should not be misrepresented as evaluating structure. The Section on 'Experimental results' below clearly indicates the evidence of the SSIM's inability to evaluate structure.

Many researches argue that an image structure is made up of edges of the objects visible in the image. By decreasing the resolution these structural features can be *washed out* [5, 6]. Most of the image structure can still be observed even if the image undergoes scale changes such as wavelet decomposition. When observed at a lower resolution, an image does not lose its structure despite losing most of the resolution. This is true for additive noise as well. Once the image is decomposed to an acceptable level, edge detection can be used to sharpen, further the structure of the image. If a metric is produced using this structural information, it will truly capture the structural information and will be a valid measure to evaluate the structural integrity, thereby making comparing images more meaningful. Our approach has been designed based on these observations.

Moment invariants based similarity measure (MISM): The moment invariants algorithm has been recognised as one of the most effective methods to extract a descriptive feature for object recognition applications and has been widely applied in the classification of subjects such as aircraft, ships, and ground targets, [7, 8]. Essentially, the algorithm derives a number of self-characteristic properties from a binary image of an object. These properties are invariant to rotation, scale and translation. Let $f(i,j)$ be a point of a digital image of size $M \times N$ ($i = 1,2, \ldots, M$ and $j = 1,2, \ldots, N$). The two dimensional moments and central moments of order $(p+q)$ of $f(i,j)$, are defined as:

$$m_{pq} = \sum_{i=1}^{M} \sum_{j=1}^{N} i^p j^q f(i,j)$$

$$U_{pq} = \sum_{i=1}^{M} \sum_{j=1}^{N} (i - \bar{i})^p (j - \bar{j})^q f(i,j)$$

where $\bar{i} = m_{10}/m_{00}$ and $\bar{j} = m_{01}/m_{00}$. From the second- and third-order moments, typically a set of seven (7) moment invariants can be derived [8]. Only the following first two moments are used for the MISM measure:

$$\varphi_1 = \eta_{20} + \eta_{02} \tag{2}$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{3}$$

where η_{pq} is the normalised central moments defined by:

$$\eta_{pq} = U_{pq}/U'_{00} \text{ for } r = [(p+q)/2] + 1, \quad p + q = 2, 3, \ldots$$

Moment invariants have been used extensively in identifying the shapes or outlines of objects for many years [7, 8]. An image reduced to 16 × 16 or larger using wavelet decomposition can be used to generate moment invariants to identify the structural make-up of an image. As our research indicates, matching at two such levels will indicate very high similarity for an image which has undergone blurring or corruption by noise and can be verified visually. Hence the approach complies with the human visual system and is far superior to MSE estimates.

An image is normalised (divided) by its own standard deviation, such that the two images being compared have unity standard deviation. An image is reduced to an approximation level (usually larger than 16 × 16) and then edges are detected using the 'Canny' operator followed by calculation of the first and second moment invariant ϕ_1 and ϕ_2 for the entire approximation [8]. Then the approximation level is divided into four quadrants and first and second moments ϕ_1 and ϕ_2 are calculated for each quadrant. These values are used to calculate the MISM for the entire image using the weights as shown in (4):

$$\mathrm{MISM} = 1 - \left(0.1 \frac{|\phi_1 - \phi'_1|}{\phi_1} + \sum_{i=1}^{4} 0.05 \frac{|\phi_{i1} - \phi'_{i1}|}{\phi_{i1}} + \sum_{i=1}^{4} 0.15 \frac{|\phi_{i2} - \phi'_{i2}|}{\phi_{i2}} \right) \tag{4}$$

Here, ϕ'_i indicates the moment invariants of the second image.

Experimental results: MISM shows a lot of promise for image similarity based metrics as well as for image matching. Fig. 1 shows a Pirate image along with a motion blurred and a noisy version. Fig. 2 shows that at deep wavelet decomposition levels, the structure is still intact. As shown in Table 1, comparing results for similarity and dissimilarity, Lena scores 0.2629 with Pirate using SSIM whereas MISM scores 0.3245. The use of Lena which has no resemblance to Pirate is used to gauge the validity of the matching process. When comparing different versions of Pirate such as Pirate with motion blur, Gaussian noise and Salt & Pepper noise, SSIM measures 0.5507, 0.3560 and 0.3831, respectively. If the SSIM truly compares structural similarity as the authors claim [3], all these images with the same structure should record a similar SSIM measure. MISM on the other hand, consistently records 0.7665, 0.9347 and 0.8367, respectively, indicating that the proposed measure is certainly measuring the structural similarity.

Fig. 1 *Images of Pirate, a motion blurred and a 0.05 Salt & Pepper noise added version of it*

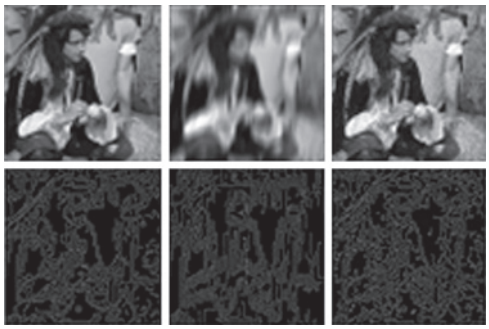

Fig. 2 *Top row: Wavelet decomposition (Daubechies 1) level 3 of Pirate, a motion blurred and a 0.05 Salt & Pepper noise added version of it. Bottom row: Respective 'Canny' edge detected versions of top*

Table 1: Comparison of SSIM and MISM for images

Image	SSIM	MISM
Pirate with motion blur	0.5507	0.7665
Pirate with 0.05 S&P noise	0.3831	0.8367
Pirate with Gaussian noise	0.3560	0.9347
Lena	0.2629	0.3245

Conclusion: We have evaluated the performance of the SSIM using the code made available by the original authors against our MISM and have demonstrated that image structural similarity can be best established accurately using MISM. In our research, we found that MISM provides more insight to the image structure as opposed to the SSIM since it does not represent structure as claimed. MISM is very much comparable to the SSIM with similar computer processing time.

18 August 2012
doi: 10.1049/el.2012.2739

P. Premaratne (*The University of Wollongong, Australia*)

E-mail: prashan@uow.edu.au

M. Premaratne (*Monash University, Australia*)

References

1 Eskicigolu, A.M.: 'Quality measurement for monochrome compressed images in the past 25 years'. Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing, Istanbul, Turkey, 2000, Vol. 4, pp. 1907–1910
2 Girod, B.: 'What's wrong with mean-squared error', Watson, A.B. (Ed.): 'Digital images and human vision' (MIT Press, 1993)
3 Wang, Z., Bovik, A., Sheik, H.R., and Simoncelli, E.P.: 'Image quality assessment: from error visibility to structural similarity', *IEEE Trans. Image Process.*, 2004, **13**, (4), pp. 1–14
4 Chen, G.H., Yang, C.L., Po, L.M., and Xie, S.L.: 'Edge-based structural similarity for image quality assessment'. Proc. Int. Conf. Acoustics Speech and Signal Processing, Toulouse, France, 2006, Vol. 2, pp. 933–936
5 Koenderink, J.J.: 'The structure of images', *Biol. Cybern.*, 1984, **50**, pp. 363–370
6 Smith, S.W.: 'The scientist's and engineer's guide to digital signal processing', *in* 'Image formation and display/digital Image structure', 1997, Chap. 23
7 Premaratne, P., Ajaz, S., and Premaratne, M.: 'Hand gesture tracking and recognition system for control of consumer electronics', *Springer Lect. Notes Artif. Intell., (LNAI)*, 2011, **6839**, pp. 588–593
8 Premaratne, P., and Nguyen, Q.: 'Consumer electronics control system based on hand gesture moment invariants', *IET Comput. Vis.*, 2007, **1**, (1), pp. 35–41