

Matched filtering determines human visual search in natural images

Alexander Toet*

TNO, P.O. Box 23, 3769 ZG Soesterberg, the Netherlands

ABSTRACT

The structural image similarity index (SSIM), introduced by Wang and Bovik (IEEE Signal Processing Letters 9-3, pp. 81-84, 2002) measures the similarity between images in terms of luminance, contrast and structure. It has successfully been deployed to model human visual perception of image distortions and modifications in a wide range of different imaging applications. Chang and Zhang (Infrared Physics & Technology 51-2, pp. 83-90, 2007) recently introduced the target structural similarity (TSSIM) clutter metric, which deploys the SSIM to quantify the similarity of a target to its background in terms of luminance, contrast and structure. They showed that the TSSIM correlates significantly with mean search time and detection probability. However, it is not immediately obvious to what extent each of the three TSSIM components contributes to this correlation. Here we evaluate the TSSIM by deploying it to a set of natural images for which human visual search data are available: the Search_2 dataset. By analyzing the predictive performance of each of the three TSSIM components, we find that it is predominantly the structural similarity component which determines human visual search performance, whereas the luminance and contrast components of the TSSIM show no relation with human performance. Since the structural similarity component of the TSSIM is equivalent to a matched filter, it appears that matched filtering predicts human visual performance when searching for a known target.

Keywords: clutter, natural images, visual search, target detection

1. INTRODUCTION

It is well known that visual targets that are similar to their local background or to details in other parts of the scene are harder to find than targets which are highly distinct. This obscuring effect, which is generally known as clutter, determines human visual search and detection performance to a large extent. Many attempts have been made to quantify the effects of clutter by means of digital clutter metrics. However, since the concept of clutter is inherently elusive, attempts to model it have only been partly successful^{1-3,6,8,17-20,24-27,29,30}.

Visual search experiments have shown that detection performance depends mainly on the energy contrast between a target and its local background, whereas recognition depends mainly on the structural dissimilarity between a target and its surround^{5,7}. For complex scenes, the spatial relationships (shape and relative location) of features in an image can have a greater effect on detection than the relative luminance of the features⁸. Higher overall contrast may even reduce the amount of perceived clutter because confusing details are more readily recognized for what they are -- nontarget scene elements. An effective clutter metric should account for this type of cognitive screening.

Wang and Bovik introduced the structural image similarity index (SSIM) which measures the similarity between images in terms of luminance, contrast and structure^{31-33,35,36}. The SSIM has successfully been deployed to model human visual perception of image distortions and modifications in a wide range of different imaging applications (for an overview see³²). Chang and Zhang^{9,10} recently introduced the TSSIM clutter metric, which deploys the SSIM to quantify the similarity of a target to its background in terms of luminance, contrast and structure. They showed that the TSSIM correlates significantly with mean search time and detection probability^{9,10}. However, it is not immediately obvious to what extent each of the three TSSIM components contributes to this correlation.

Here we analyze the predictive performance of each of the three TSSIM components, and we show that it is predominantly the structural similarity component which determines human visual search performance, whereas the luminance and contrast components of the TSSIM do not correlate with human performance. The rest of this paper is organized as follows. In Section 2 we show how rewriting the TSSIM in its full form allows the assessment of the contribution of the luminance, contrast and structural similarity components to the overall clutter metric. In Section 3 we

describe how the performance of the TSSIM was evaluated by deployment to a set of natural images for which human observer data are available. The results of this experiment are presented in Section 4. Finally, the conclusions of this study are presented in Section 5.

2. CLUTTER METRICS

2.1 The structural similarity (SSIM) metric

Let $x = \{x_i \mid i = 1, 2, \dots, N\}$ and $y = \{y_i \mid i = 1, 2, \dots, N\}$ represent two discretely sampled grayscale image patches that need to be compared. Let $\mu_x, \mu_y, \sigma_x, \sigma_y, \sigma_{xy}$ respectively be the mean of x , the mean of y , the standard deviation of x , the standard deviation of y , and the covariance of x and y , defined as:

$$\mu_x = \frac{1}{N} \sum_{i=1}^N x_i, \quad \mu_y = \frac{1}{N} \sum_{i=1}^N y_i \quad (1)$$

$$\sigma_x = \left(\frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_x)^2 \right)^{\frac{1}{2}}, \quad \sigma_y = \left(\frac{1}{N-1} \sum_{i=1}^N (y_i - \mu_y)^2 \right)^{\frac{1}{2}} \quad (2)$$

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y) \quad (3)$$

The mean signal intensity and its standard deviation (the square root of variance) can be regarded as rough estimates of respectively local image luminance and contrast. The covariance of x and y (normalized by their respective variances) reflects the tendency of the two signals to vary together, and can therefore be adopted as a measure of the structural similarity between the two signals.

The similarity of the local patch luminances is then defined as

$$l(x, y) = \frac{2\mu_x \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (4)$$

where C_1 is a small constant given by $C_1 = (K_1 L)^2$, which is introduced to stabilize the computation of (4) when the denominator becomes small, L is the dynamic range of the pixel values ($L = 255$ for 8 bits/pixel grayscale images), and $K \ll 1$ is a scalar constant (typically 0.01). The dynamic range of l is $\langle 0, 1 \rangle$. The maximum value 1 is approached when both image patches have the same luminance: $\mu_x = \mu_y$.

The similarity of the local patch contrasts is defined as

$$c(x, y) = \frac{2\sigma_x \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (5)$$

where C_2 is a small constant given by $C_2 = (K_2 L)^2$, $K_2 \ll 1$ (typically 0.03). The dynamic range of σ is $\langle 0, 1 \rangle$. The maximum value 1 is approached when both image patches have the same contrast: $\sigma_x = \sigma_y$.

The structural similarity between the image patches is defined as

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3} \quad (6)$$

with $C_3 = C_2/2$. The dynamic range of s is $\langle -1, 1 \rangle$. The maximum value 1 is approached when $y_i = ax_i + b$ for all $i = 1, 2, \dots, N$, where a and b are constants and $a > 0$.

The overall structural similarity index SSIM between signals x and y is defined as:

$$\text{SSIM}(x, y) = |l(x, y)|^\alpha \cdot |c(x, y)|^\beta \cdot |s(x, y)|^\gamma \quad (7)$$

where α, β and γ are parameters that define the relative importance of the three components.

Setting $\alpha = \beta = \gamma = 1$ and substitution of Equations (4),(5) and (6) in Equation (7) results in

$$\text{SSIM}(x, y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (8)$$

which is the form in which the SSIM is typically used in the literature³²⁻³⁶.

When comparing two images, the SSIM index is computed locally within a sliding window that moves pixel-by-pixel across the image, resulting in a SSIM map. The SSIM score of the entire image is then computed by pooling the SSIM map, e.g., by simply averaging the SSIM values across the image. SSIM has successfully been applied in a large number of different applications (for an overview see³²).

2.2 The target structural similarity (TSSIM) clutter metric

Chang and Zhang^{9,10} adapted the SSIM for use as a clutter metric, and introduced the target structure similarity metric (TSSIM). Their approach is as follows:

The image for which the clutter metric has to be calculated is divided into N blocks. The blocks are twice the apparent size of the typical search target in each dimension.

Let $T = \{T_i \mid i = 1, 2, \dots, N\}$ and $B_j = \{B_{ji} \mid i = 1, 2, \dots, N\}$ represent respectively the discretely sampled grayscale target block (i.e. the part of the images which contains the target support area and a local background area around the target) and the j^{th} background image block.

Substitution of T and B for x and y in Equations (4)-(6) and neglecting the stabilizing constants yields

$$l(T, B_j) = \frac{2\mu_T \mu_{B_j}}{\mu_T^2 + \mu_{B_j}^2} \quad (9)$$

$$c(T, B_j) = \frac{2\sigma_T \sigma_{B_j}}{\sigma_T^2 + \sigma_{B_j}^2} \quad (10)$$

$$s(T, B_j) = \frac{\sigma_{TB_j}}{\sigma_T \sigma_{B_j}} \quad (11)$$

Substituting Equations (9), (10) and (11) in (7), with $\alpha = \beta = \gamma = 1$, and adopting a single constant C to avoid instabilities, yields the TSSIM metric^{9,10}:

$$\text{TSSIM}(T, B_j) = \frac{4\mu_T \mu_{B_j} \sigma_{TB_j} + C}{(\mu_T^2 + \mu_{B_j}^2)(\sigma_T^2 + \sigma_{B_j}^2) + C} \quad (12)$$

Chang and Zhang used both $C = 0.2$ ⁹ or $C = 0$ ¹⁰. Here we also adopt $C = 0$ since we observed no instabilities for the image set used in our experiments.

The overall image TSSIM is then calculated in two ways: both as the root mean square of TSSIM_j (TSSIM_{rms}) and the arithmetic mean of TSSIM_j (TSSIM_{am}).

$$\text{TSSIM}_{rms} = \sqrt{\frac{1}{N} \sum_{j=1}^N \text{TSSIM}_j^2} \quad (13)$$

$$\text{TSSIM}_{am} = \frac{1}{N} \sum_{j=1}^N \text{TSSIM}_j \quad (14)$$

where TSSIM_j is the structure similarity measure in the j^{th} image block, and N is the total number of image blocks. The rationale of this metric is the fact that observers will need more time to inspect the image when it contains more details similar to the target. Details similar to the search target can also distract and confuse the observer, and may result in false alarms, thus degrading the detection probability. Thus, a higher TSSIM value corresponds to more clutter in the image, leading to longer search (inspection) times and lower detection probability.

2.3 The full form TSSIM

Using Equations (9), (10) and (11), and setting all stabilization constants to zero, we rewrite Equation (12) in its original full form (Equation (7)), which allows the assessment of the individual contributions of luminance, contrast and structural similarity to the overall TSSIM clutter metric:

$$\text{TSSIM}(T, B_j) = l(T, B_j) \cdot c(T, B_j) \cdot s(T, B_j) = \frac{2\mu_T \mu_{B_j}}{\mu_T^2 + \mu_{B_j}^2} \cdot \frac{2\sigma_T \sigma_{B_j}}{\sigma_T^2 + \sigma_{B_j}^2} \cdot \frac{\sigma_{TB_j}}{\sigma_T \sigma_{B_j}} \quad (15)$$

In the next section we will use Equations (9), (10), (11), and (15) to investigate the contribution of each of the perceptually relevant factors (i.e. the luminance component $l(T, B_j)$, the contrast component $c(T, B_j)$, and structural similarity component $s(T, B_j)$) to the overall TSSIM clutter measure.

3. EXPERIMENT

Similar to Chang and Zhang^{9,10}, we use the Search_2 image dataset²⁸ to assess the performance of the TSSIM (Equation (15)) and each of its three components (9), (10) and (11). The Search_2 dataset consists of a set of 44 high-resolution digital color images of different complex natural scenes including a search target (a military vehicle), together with the detection times and detection probability obtained from a visual search and detection experiment in which 62 observers participated. The Search_2 dataset, which is described in detail elsewhere²⁸, has been extensively used in different studies in the literature, ranging from studies evaluating target detectability and clutter metrics, to eye movement studies and attempts to model the human visual system^{4,10-16,37}. Similar to Chang and Zhang^{9,10}, we converted the original Search_2 images to grayscale and reduced their resolution with a factor 2, resulting in an image size of 3072x2048 pixels. We used the same target blocks as Chang and Zhang^{9,10}. In addition to the TSSIM, we also computed its luminance component l (Equation (9)), its contrast component c (Equation (10)), and its structure component s (Equation (11)), for all 44 Search_2 images.



(a)



(b)

Fig. 1 Two images from the Search_2 database, divided into blocks (white rectangles) with twice the apparent size of the search target. The target blocks are represented by the red rectangles.

4. RESULTS

Table 1 and Table 2 show respectively the correlation between the TSSIM and each of its three components l , c and s on the one hand, and the logarithm of the mean search time and the detection probabilities for the images in the Search_2 dataset on the other hand, quantified by respectively the Pearson and the Spearman rank order correlation coefficients. The correlations were computed using SPSS 18.0. These results show that the TSSIM and its structural similarity component s correlate significantly at the 0.01 level (1-tailed) with the logarithm of the mean search time and the detection probability, whereas the luminance and contrast similarity components show no significant correlation. TSSIM and s both correlate most strongly with the logarithm of the mean search time. The structure component s shows a stronger correlation with both Pd and logST than TSSIM. The largest correlation values are obtained for the arithmetic mean of s . The overall largest correlation is obtained between the arithmetic mean of s and logST (Pearson correlation = .812, Spearman rank order correlation = .826).

Note that the structural similarity component s is in fact a correlation measure, similar to a matched filter. The TSSIM parses the image into non-overlapping blocks, performs "matched filtering" on each block, and computes the overall clutter metric as the (arithmetic or root mean square) average over all blocks. The rationale for this approach is that the mean search time will be longer, and the detection probability will be smaller, if there are more image blocks that "match" (are similar to) the search target. To investigate the effects of the use of discrete blocks, we computed the TSSIM for blocks with different degrees of overlap, ranging from zero overlap (non-overlapping image tessellation, as used by the TSSIM) to maximal overlap (corresponding to full matched filtering). We found that all (Pearson and Spearman) correlations thus obtained vary less than 7%.

Table 1. The correlation between TSSIM and each of its three components l , c and s , and the logarithm of the mean search time and the detection probabilities for the images in the Search_2 dataset, quantified by the Pearson correlation coefficient. Bold values indicate that the correlation is significant at the 0.01 level (1-tailed).

	l_{rms}	l_{am}	c_{rms}	c_{am}	s_{rms}	s_{am}	TSSIM _{rms}	TSSIM _{am}
logST	-.046	-.048	-.187	-.177	.807	.812	.789	.781
Pd	.049	.050	.087	.081	-.641	-.647	-.639	-.636

Table 2. The correlation between TSSIM and each of its three components l , c and s , and the logarithm of the mean search time and the detection probabilities for the images in the Search_2 dataset, quantified by the Spearman rank order correlation coefficient. Bold values indicate that the correlation is significant at the 0.01 level (1-tailed).

	l_{rms}	l_{am}	c_{rms}	c_{am}	s_{rms}	s_{am}	TSSIM _{rms}	TSSIM _{am}
logST	.005	.012	-.239	-.213	.816	.826	.767	.755
Pd	-.047	-.058	.137	.112	-.759	-.764	-.744	-.760

5. CONCLUSIONS

In this study we first replicated the results of Chang and Zhang^{9,10}, who showed that the newly introduced target structural similarity clutter metric TSSIM correlates significantly with human visual search performance, and outperforms other clutter metrics. Then we showed that it is the structural similarity component of the TSSIM that predicts both mean search time and detection probability, while the luminance and contrast similarity components do not correlate with human observer performance. This result agrees with the related observations that the cross-correlation component of the SSIM predicts visual image quality²³, and that human observers mainly rely on structural features to recognize image content^{5,7,8,21,22}. Furthermore, it should be noted that the structural similarity component of the TSSIM is equivalent to a matched filter. Hence, it appears that matched filtering predicts human visual performance when searching for a known target. However, since the image dataset used in this study is limited, further experiments on a wide variety of images (preferably of different modalities) are required to establish the general validity of the structural similarity component as a clutter metric.

Acknowledgement

The author thanks dr Honghua Chang for kindly providing all the information and assistance needed to reproduce his previous results.

REFERENCES

1. Aviram, G. and Rotman, S.R., Evaluating human detection performance of targets and false alarms, using a statistical texture image metric, *Optical Engineering*, 39(8) ,pp. 2285-2295, 2000.
2. Aviram, G. and Rotman, S.R., Evaluating TNO human target detection experimental results agreement with various image metrics, In: A. Toet (Ed.), *Search and Target Acquisition*, pp. 1-6, North Atlantic Treaty Organization, Neuilly-sur-Seine Cedex, France, 2000.
3. Aviram, G. and Rotman, S.R., Evaluation of human detection performance of targets embedded in natural and enhanced infrared images using image metrics, *Optical Engineering*, 39(4) ,pp. 885-896, 2000.
4. Birkemark, C.M., CAMEVA, a methodology for estimation of target detectability, *Optical Engineering*, 40(9) ,pp. 1835-1843, 2001.
5. Braje, W.L., Tjan, B.S. and Legge, G.E., Human efficiency for recognizing and detecting low-pass filtered objects, *Vision Research*, 35(21) ,pp. 2955-2966, 1995.
6. Bravo, M.J. and Farid, H., A scale invariant measure of clutter, *Journal of Vision*, 8(1) ,pp. 23-1-23-9, 2008.
7. Caelli, T. and Moraglia, G., On the detection of signals embedded in natural scenes, *Perception & Psychophysics*, 39(2) ,pp. 87-95, 1986.
8. Cathcart, J.M., Doll, T.J. and Schmieder, D.E., Target detection in urban clutter, *IEEE Transactions on Systems, Man and Cybernetics SMC*, 19(5) ,pp. 1242-1250, 1989.
9. Chang, H. and Zhang, J., New metrics for clutter affecting human target acquisition, *IEEE Transactions on Aerospace and Electronic Systems*, 42(1) ,pp. 361-368, 2006.
10. Chang, H. and Zhang, J., Detection probability and detection time using clutter metrics, *Infrared Physics & Technology*, 51(2) ,pp. 83-90, 2007.
11. Garcia, J.A., Fdez-Valdivia, J., Fdez-Vidal, X.R. and Rodriguez-Sánchez, R., Information theoretic measure for visual target distinctness, *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 23(4) ,pp. 362-383, 2001.
12. Garcia, J.A., Fdez-Valdivia, J., Fdez-Vidal, X.R., Rodriguez-Sánchez, R. and Fuertes, J.M., Minimum error gain for predicting visual target distinctness, *Optical Engineering*, 40(9) ,pp. 1794-1817, 2001.
13. Garcia-Díaz, A., Fdez-Vidal, X.R., Pardo, X.M. and Dosil, R., Local energy variability as a generic measure of bottom-up salience, In: P.-Y. Yin (Ed.), *Pattern Recognition Techniques, Technology and Applications*, pp. 626-650, I-Tech, Vienna, Austria, 2008.

14. Itti, L., Gold, C. and Koch, C., Visual attention and target detection in cluttered natural scenes , *Optical Engineering*, 40(9) ,pp. 1784-1793, 2001.
15. Meitzler, T.J., Sohn, E., Singh, H. and Elgarhi, A., Predicting search time in visual scenes using the fuzzy logic approach, *Optical Engineering*, 40(9) ,pp. 1844-1851, 2001.
16. Nilsson, T., Evaluation of target acquisition difficulty using recognition distance to measure required retinal area, *Optical Engineering*, 40(9) ,pp. 1827-1834, 2001.
17. Rosenholtz, R., Li, Y., Mansfield, J. and Jin, Z., Feature congestion: a measure of display clutter, In: *Proceeding of the SIGCHI conference on Human factors in computing systems*, pp. 761-770, ACM Press, New York, USA, 2005.
18. Rosenholtz, R., Li, Y. and Nakano, T., Measuring visual clutter, *Journal of Vision*, 7(2) ,pp. 71-1-71-22, 2007.
19. Rotman, S.R., Hsu, D., Cohen, A., Shamay, D. and Kowalczyk, M.L., Textural metrics for clutter affecting human target acquisition, *Infrared Physics and Technology*, 37(6) ,pp. 667-674, 1996.
20. Rotman, S.R., Tidhar, G. and Kowalczyk, M.L., Clutter metrics for target detection systems, *IEEE Transactions on Aerospace and Electronic Systems*, 30(1) ,pp. 81-91, 1994.
21. Rouse, D.M. and Hemami, S.S., Analyzing the role of visual structure in the recognition of natural image content with multi-scale SSIM, In: *Proceedings of the IEEE Western New York Image Processing Workshop (WNYIP)*, pp. Rochester, NY, 2007.
22. Rouse, D.M. and Hemami, S.S., Quantifying the use of structure in cognitive tasks, In: B.E. Rogowitz, T.N. Pappas & S.J. Daly (Ed.), *Human Vision and Electronic Imaging XII*, pp. 1-10, Society of Photo-Optical Instrumentation Engineers, Bellingham, WA, 2007.
23. Rouse, D.M. and Hemami, S.S., Understanding and simplifying the structural similarity metric, In: *Proceedings of the 15th IEEE International Conference of Image Processing (ICIP2008)*, pp. 1188-1191, IEEE Press, 2008.
24. Salem, S., Halford, C., Moyer, S. and Gundy, M., Rotational clutter metric, *Optical Engineering*, 48(086401) ,pp. 1-11, 2009.
25. Schmieder, D.E. and Weathersby, M.R., Detection performance in clutter with variable resolution, *IEEE Transactions on Aerospace and Electronic Systems*, 19(4) ,pp. 622-630, 1983.
26. Shirvaikar, M.V. and Trivedi, M.M., Developing texture-based image clutter measures for object detection, *Optical Engineering*, 31 ,pp. 2628-2639, 1992.
27. Tidhar, G., Reiter, G., Avital, Z., Hadar, Y., Rotman, S.R., George, V. and Kowalczyk, M.L., Modeling human search and target acquisition performance: IV. detection probability in the cluttered environment, *Optical Engineering*, 33 ,pp. 801-808, 1994.
28. Toet, A., Bijl, P. and Valetton, J.M., Image dataset for testing search and detection models, *Optical Engineering*, 40(9) ,pp. 1760-1767, 2001.
29. van den Berg, R.V., Cornelissen, F.W. and Roerdink, J.B.T.M., A crowding model of visual clutter, *Journal of Vision*, 9(4) ,pp. 24-1-24-11, 2009.
30. Waldman, G., Wootton, J., Hobson, G. and Luetkemeyer, K., A normalized clutter measure for images, *Computer Vision, Graphics and Image Processing*, 42 ,pp. 137-156, 1988.
31. Wang, Z. and Bovik, A.C., A universal image quality index, *IEEE Signal Processing Letters*, 9(3) ,pp. 81-84, 2002.
32. Wang, Z. and Bovik, A.C., Mean squared error: love it or leave it? - A new look at signal fidelity measures, *IEEE Signal Processing Magazine*, 26(1) ,pp. 98-117, 2009.
33. Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P., Image quality assessment: from error visibility to structural similarity, *IEEE Transactions on Image Processing*, 13(4) ,pp. 600-612, 2004.
34. Wang, Z., Lu, L. and Bovik, A.C., Video quality assessment based on structural distortion measurement, *Signal Processing: Image Communication*, 19(2) ,pp. 121-132, 2004.
35. Wang, Z., Sheikh, H.R., and Bovik, A.C. (2003). Objective video quality assessment. In: B. Furht & O. Marqure (Eds.), *The Handbook of Video Databases: Design and Applications*. (pp. 1041-1078). Boca Raton, Florida: CRC Press.
36. Wang, Z., Simoncelli, E.P. and Bovik, A.C., Multi-scale structural similarity for image quality assessment, In: *Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems and Computers*, pp. 1398-1402, 2003.
37. Yang, C., Zhang, J., Xu, X., Chang, H.-H. and He, G.-J., Quaternion phase-correlation-based clutter metric for color images, *Optical Engineering*, 46(12) ,pp. 127008-1-127008-7, 2007.