STRUCTURAL TEXTURE SIMILARITY METRIC BASED ON INTRA-CLASS VARIANCES

Matteo Maggioni[†]*, Guoxin Jin*[⋆]*, Alessandro Foi*[†]*, Thrasyvoulos N. Pappas*[⋆]

[†]Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[⋆]EECS Department, Northwestern University, Evanston, IL, USA

ABSTRACT

Traditional point-by-point image similarity metrics, such as the ℓ_2-norm, are not always consistent with human perception, especially in textured regions. We consider the problem of identifying textures that are perceptually identical to a query texture; this is important for image retrieval, compression, and restoration applications. Recently proposed structural texture similarity (STSIM) metrics assign high similarity scores to such perceptually identical textured patches, even though they may have significant pixel-wise deviations. We use an STSIM approach that compares a set of statistical patch descriptors through a weighted distance, and, given a dataset of labeled texture images partitioned into classes of perceptually identical patches, we calculate the weights as the variances of each statistic centered around the mean of its class. Experimental results demonstrate that the proposed approach outperforms existing structural similarity metrics and STSIMs as well as traditional point-by-point metrics when assessing texture similarity in both noisy and noise-free conditions.

Index Terms— Perceptual equivalence, structural similarity, statistical analysis, content-based retrieval

1. INTRODUCTION

In the past decade, image redundancy and self-similarity enabled a substantial improvement for several image processing applications, and, although such redundancy is used in different forms by different algorithms, determining image similarity is always a task of fundamental interest [1–4].

The traditional strategy for measuring patch similarity is based on point-by-point metrics, such as the Euclidean distance (ℓ_2-norm) of the difference of the data being compared. However, such metrics are not always consistent with human perception, especially when comparing textured content, because they fail to account for the typically stochastic characteristics of natural textures. Thus, there is an increasing interest in structural texture similarity (STSIM) metrics that replace point-by-point comparisons with statistical approaches that compare a set of patch statistics [4–6]. We focus on the retrieval of "perceptually identical" textures (henceforth *identical*) such as those in Fig. 1, which can be considered to be pieces of a larger perceptually uniform texture [6]. Distinguishing identical and nonidentical textures is important for a variety of applications including content-based retrieval, restoration, and compression [4, 6–8]. STSIM metrics are capable of assigning high similarity values to patches with significant pixel-wise deviations that are perceived as essentially identical.

In this paper we adopt the STSIM-M formulation [6], in which the statistics of two patches are compared using a Mahalanobis distance [9]. In particular, given a dataset of labeled texture images partitioned into classes of identical patches, we propose a variation of the STSIM-M metric, whereby the covariance matrix of the Mahalanobis distance is diagonal (as in STSIM-M), but each diagonal entry is computed as the variance of the corresponding statistic centered around the mean of its class, thus capturing the intra-class variance (also known in literature as "within-class" variance [10]) instead of the global variance as is done in STSIM-M.

We evaluate the proposed metric in the context of the retrieval of identical textures using a labeled database that consists of noisy or noise-free 1181 grayscale patches divided into 425 classes [6]. For our purposes, the noise should have no structure, so we consider an i.i.d. additive Gaussian white noise. Our results demonstrate that, in all cases, the proposed strategy outperforms structural similarity metrics (SSIMs) [11, 12] and previous manifestations of STSIMs [6].

The remainder of the paper is organized as follows. Section 2 reviews STSIMs, setting the stage for the formulation of the new metric framework, which is presented in Section 3. The experimental results are reported in Section 4 and the conclusions are summarized in Section 5.

2. REVIEW OF STRUCTURAL TEXTURE SIMILARITY METRICS

The development of STSIMs was inspired by the introduction of the structural similarity metric (SSIM) [11], and its transform domain (complex wavelet) extension CW-SSIM [12]. However, even though they represent an attempt to move away from point-by-point comparisons, SSIM and CW-SSIM

♡ Work supported by Academy of Finland (project no. 252547, Academy Research Fellow 2011-2016), Tampere Graduate School in Information Science and Engineering (TISE), KAUTE foundation, and Nokia foundation.

still incorporate a point-by-point comparison term (the "structure" term). STSIMs [5,6] on the other hand, rely completely on patch statistics, and are characterized by the following main stages:

1. *Subband Decomposition.* The compared patches undergo a multi-scale multi-oriented frequency decomposition, such as the steerable filter [13], to mimic the biological processing of the human visual system.

2. *Extraction of Statistics.* Each patch is represented by a number of statistical descriptors separately computed over the individual subbands.

3. *Statistics Comparison.* A final similarity score is given by defining a strategy to compare the statistics of each patch.

In [6], Zujovic *et al.* presented a number of STSIM formulations, including the one proposed initially by Zhao *et al.* [5]. In this paper, we propose a variation of the STSIM-M formulation [6]. Let z_i and z_j be two image patches and let $\omega_{i,k}$ and $\omega_{j,k}$ with $k = 1, \ldots, m$, be independent statistics extracted from z_i and z_j, respectively [6]. Then, the STSIM-M between two patches z_i and z_j can be defined as the following particular case of Mahalanobis distance [9]

$$\sqrt{\sum_{k=1}^{m} \frac{\left(\omega_{i,k} - \omega_{j,k}\right)^2}{\sigma_k^2}}, \tag{1}$$

where σ_k^2 is the variance of the k-th statistic over a training set of patches $\Gamma = \{z_i : i = 1, \ldots, n\}$ defined as

$$\sigma_k^2 = \text{var}\{\omega_{i,k}\} = \frac{1}{n-1} \sum_{i=1}^{n} \Big(\omega_{i,k} - \text{mean}\{\omega_{i,k}\}\Big)^2 \tag{2}$$

where $\text{mean}\{\omega_{i,k}\} = 1/n \sum_{i=1}^{n} \omega_{i,k}$.

The development of the STSIM metrics was motivated by the texture analysis/synthesis model proposed in [14], which in its entirety contains 846 parameters; Zujovic *et al.* select a subset of $m = 82$ statistics for (1) which are considered sufficient for the development of a STSIM [6].

3. INTER-CLASS AND INTRA-CLASS VARIANCES

The STSIM-M formulation allows to differently weight different statistics. In particular, statistics with larger variance are penalized by (1) as they are expected to be less informative for the overall metric. However, the problem with this approach is that it does not discriminate between the variance due to the inherent statistic fluctuation (within a perceptually uniform texture) and the variance due to fluctuations of the statistic that are due to differences in content (across different textures). Thus, we identify two additive terms within the overall variance σ_k^2 of the k-th statistic:

$$\sigma_k^2 = \varsigma_k^2 + \vartheta_k^2 \tag{3}$$

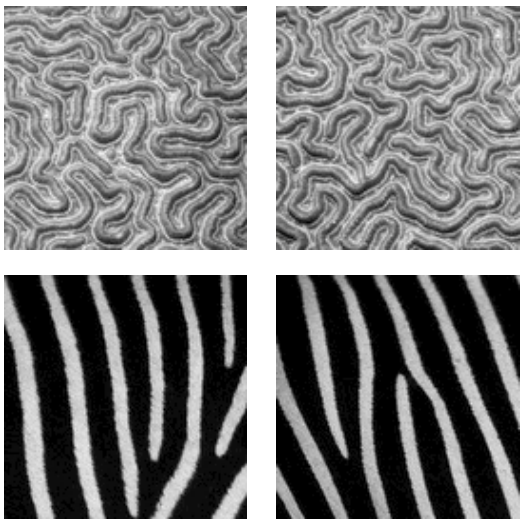

Fig. 1. Example of perceptually identical patches.

where the first term ς_k^2 is the inherent (intra-class) variance of the k-th statistic, while the second term ϑ_k^2 accounts for the variation of the k-th statistic across dissimilar patches.

Formally, let $\Lambda = \{\lambda_i : i = 1, \ldots, n\}$ be the set of n labels $\lambda_i \in \mathbb{N}^+$ for the patches in the training set Γ. Two labels λ_i and λ_j are equal if and only if the corresponding patches z_i and z_j are perceptually identical. In this way we can explicitly define the class of (indices of) patches identical to any given patch $z_i \in \Gamma$ as

$$\Theta(z_i) = \big\{j : \lambda_i = \lambda_j, j = 1, \ldots, n\big\}. \tag{4}$$

As a reference, we show in Fig. 1 two perceptually identical patches. Note that, despite having a significant point-by-point difference, the two patches looks identical according to human perception.

Then, from each $\Theta(z_i)$, we obtain the class-wise mean of the k-th statistic as

$$\mu_{i,k} = \frac{1}{|\Theta(z_i)|} \sum_{j \in \Theta(z_i)} w_{j,k}, \tag{5}$$

where $|\Theta(z_i)|$ is the cardinality of $\Theta(z_i)$. Observe that, whenever $\lambda_i = \lambda_j$ then we also have $\mu_{i,k} = \mu_{j,k}$. With (5), we calculate the intra-variance ς_k^2 as

$$\varsigma_k^2 = \text{var}\big\{\omega_{i,k} - \mu_{i,k}\big\} = \frac{1}{n-1} \sum_{i=1}^{n} \big(\omega_{i,k} - \mu_{i,k}\big)^2, \tag{6}$$

because by construction $\text{mean}\{\omega_{i,k} - \mu_{i,k}\} = 0$. Intuitively, (6) is the sample variance of the k-th statistic of every patch $z_i \in \Gamma$ centered around its class mean (5) discarding the influence of the fluctuations due to inter-class variability, which is instead captured by the second term ϑ_k^2. Note that the inter-class variance ϑ_k^2 can be easily obtained by subtracting (6) from (2), or as $\vartheta_k^2 = \text{var}\{\mu_{i,k}\}$.

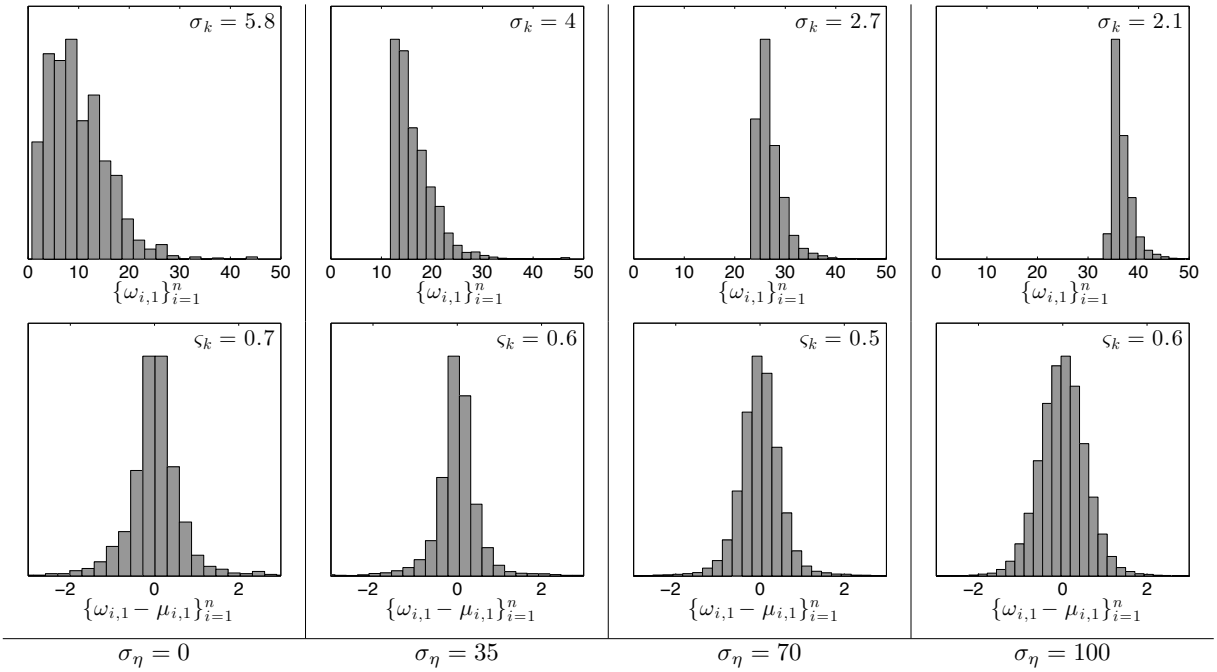

Fig. 2. Distributions of the statistic $\omega_{i,1}$ (i.e. the mean of the modulus of the first subband in the decomposition [13]) for all the patches in the image database [6] before (top row) and after (bottom row) the subtraction of (5) under different noise levels σ_η.

Since the two terms composing (3) are available, we can define a structural similarity metric between patches z_i and z_j accounting for the inter- and intra-class variances as

$$\sqrt{\sum_{k=1}^{m} \frac{\left(\omega_{i,k}/\vartheta_k - \omega_{j,k}/\vartheta_k\right)^2}{\varsigma_k^2/\vartheta_k^2}} \equiv \sqrt{\sum_{k=1}^{m} \frac{\left(\omega_{i,k} - \omega_{j,k}\right)^2}{\varsigma_k^2}}, \quad (7)$$

where the statistics $\omega_{\cdot,k}$ are first divided by the inter-class standard deviation ϑ_k to equalize their response to inter-class differences, and then by their intra-class variance ς_k divided by ϑ_k which acts as a signal-to-noise ratio between the quantity of interest ς_k and its dispersion ϑ_k. The metric (7) emphasize differences between statistics whose intra-class variance (6) is small. The advantage of using (6) over (3) is that (6) is a better representation of the actual variability of the statistics. In particular, our model evaluates the similarity of the patches by sphering the data through the intra-class variances around the centers of mass (5) of the classes.

4. EXPERIMENTS

In this section we evaluate the performance of the proposed metric (7), which we denote STSIM-I, with a series of image retrieval tests. We use a database of 1181 grayscale noise-free patches originating from 425 high-resolution images characterized by perceptually uniform textures[1]. The extracted patches have size 128×128 pixel, and patches extracted from the same source image are considered to belong to the same class, i.e. to be perceptually identical. For the noisy case, we generate noisy patches z_i following the standard observation model

$$z_i = y_i + \eta_i, \quad (8)$$

where y_i is the noise-free patch, and $\eta_i \sim \mathcal{N}(0, \sigma_\eta^2)$ is i.i.d. Gaussian noise with variance σ_η^2. All images are hereafter considered to be in the range $[0, 255]$.

For the subband decomposition, we use the steerable filter [13] using four orientations and three scales without subsampling. Thus obtaining a pyramid of twelve complex subbands plus one real innermost lowpass band and one real outermost highpass band. Every set $\Omega(z_i)$ contains $m = 82$ statistics: mean, variance, and vertical and horizontal autocorrelation for each subbands, and one cross-correlation for each pair of subbands at a given scale and for each pair of subbands adjacent in scale. The statistics are calculated from the magnitude of the subband coefficients as it has been proven to be more effective [6]. In Fig. 2, we show the distribution of the statistics $w_{i,1}$ (i.e. the mean of the modulus of the first subband) of all the patches in the database before and after the subtraction of the class means (5) under the different noise conditions.

[1]The dataset was originally designed in [6] using images from the Corbis database http://www.corbisimages.com/

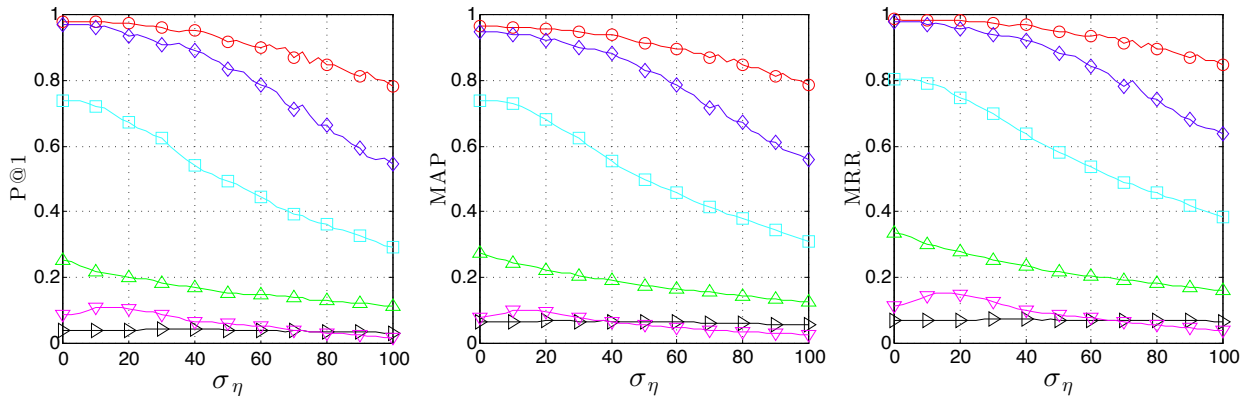

Fig. 3. From left to right: precision at one (P@1), mean average precision (MAP) and mean reciprocal rank (MRR) of STSIM-I (red ○), STSIM-M (blue ⋄), STSIM-2 (cyan □), SSIM (magenta ▽), CW-SSIM (green △), and MSE (black ▷).

We compare the retrieval performance of the proposed STSIM-I and STSIM-M [6] against SSIM [11], CW-SSIM [12], STSIM-2 [6], as well as MSE. For every noise level $\sigma_\eta \in \{0, 2.5, 5, \ldots, 100\}$, we measure the similarity between a given patch (the query) and the rest of the patches (the targets) in the database, and then we sort the results in descending value of similarity. Since Λ (i.e. the class of each patch) is known, we can measure the performance of the compared metrics using three criteria:

- *Precision at one* (P@1): percentage of cases for which the target with the highest similarity score is identical to the query;
- *Mean Average Precision* (MAP): average goodness in the ordering of the targets [15];
- *Mean Reciprocal Rank* (MRR): average distance of the first target identical to the query.

For both STSIM-M and STSIM-I we perform a K-fold cross-validation with $K = 5$. In every K run, we extract a different training set Γ comprising of 85 classes and n_K patches from our original image database. The remaining patches will be used as validation data. In order to increase the number of training samples, we generate $M = 100$ Montecarlo realizations of every training patch in Γ following (8) for each considered noise level σ_η. Then, we calculate independently for each σ_η the variances (6) and (3) from the m statistics of all $M \cdot n_K$ training patches. The final performances of STSIM-M and STSIM-I are obtained by averaging the results of the K cross-validations.

In Fig. 3, we report the performances of STSIM-I (red ○), STSIM-M (blue ⋄), STSIM-2 (cyan □), SSIM (magenta ▽), CW-SSIM (green △), and MSE (black ▷) according to P@1, MAP, and MRR as a function of the σ_η. The metrics STSIM-2 and CW-SSIM also embed the decomposition [13], thus we use the same scale and orientation parameters of STSIM-I and STSIM-M. All other parameters are set to their default values. As one can clearly see, the structural similarity metrics STSIM-I, STSIM-M and STSIM-2 clearly outperforms SSIM, CW-SSIM and MSE, and, in particular, the proposed STSIM-I achieves an improvement over STSIM-M that ranges between 2% (in the noise-free case) and 30% (in the most noisy case $\sigma_\eta = 100$).

The performance degradation of STSIM-I as σ_η increases is also more graceful than that of STSIM-M, thus demonstrating a superior ability to discern unstructured noise from stochastic texture.

5. CONCLUSIONS

We presented a structural similarity metric, called STSIM-I, that accounts for the stochastic nature of patches and is able to cope with the presence of Gaussian noise. The metric, inspired by [6], identifies patches that according to human judgment are perceptually identical even if there exists a substantial point-by-point difference. The metric is based on the filter [13], which provides a multi-scale multi-orientated subband decomposition of the patch, and on a set of statistics calculated over such subbands. Then, the statistics of two patches are compared using a weighted distance. We propose to calculate the weights as the intra-class variance of each statistic centered around the mean of the class of perceptually identical patches.

Our experiments demonstrate that the proposed approach outperforms the structural metrics STSIM-M, STSIM-2, SSIM and CW-SSIM, as well as the traditional MSE in both noisy and noise-free condition. Future work addresses a direct applicability of the proposed metric as well as the statistical representation of patches in the restoration/denoising problem.

6. REFERENCES

[1] D. L. Ruderman, "The statistics of natural images," *Network: Computation in Neural Systems*, vol. 5, no. 4, pp. 517–548, Nov 1994.

[2] V. Katkovnik, A. Foi, K. Egiazarian, and J. Astola, "From local kernel to nonlocal multiple-model image denoising," *International Journal of Computer Vision*, vol. 86, pp. 1–32, Jan. 2010.

[3] P. Milanfar, "A tour of modern image filtering: New insights and methods, both practical and theoretical," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 106–128, Jan. 2013.

[4] T.N. Pappas, D.L. Neuhoff, H. de Ridder, and J. Zujovic, "Image analysis: Focus on texture similarity," *Proc. of the IEEE*, vol. 101, no. 9, pp. 2044–2057, Sep. 2013.

[5] X. Zhao, M.G. Reyes, T.N. Pappas, and D.L. Neuhoff, "Structural texture similarity metrics for retrieval applications," in *IEEE International Conference on Image Processing*, Oct 2008, pp. 1196–1199.

[6] J. Zujovic, T.N. Pappas, and D.L. Neuhoff, "Structural texture similarity metrics for image analysis and retrieval," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2545–2558, Jul. 2013.

[7] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3D transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[8] G. Jin, T.N. Pappas, and D.L. Neuhoff, "An adaptive lighting correction method for matched-texture coding," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2014.

[9] P. C. Mahalanobis, "On the generalized distance in statistics," *Proc. of the National Institute of Sciences of India*, vol. 2, pp. 49–55, Jan. 1936.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[11] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[12] Z. Wang and E.P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Mar. 2005, vol. 2, pp. 573–576.

[13] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, and D.J. Heeger, "Shiftable multiscale transforms," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 587–607, Mar. 1992.

[14] J. Portilla and E.P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 49–70, 2000.

[15] E.M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," *Information Processing & Management*, vol. 36, no. 5, pp. 697–716, 2000.