Visual and data stationarity of texture images

Michele Conni,[a,b,*] **Hilda Deborah**,[a] **Peter Nussbaum**,[a] **and Phil Green**[a]
[a]Norwegian University of Science and Technology, Department of Computer Science,
Gjøvik, Norway
[b]Barbieri Electronic, Brixen, Italy

Abstract. The stationarity of a texture can be considered a fundamental property of images, although the property of stationarity is difficult to define precisely. We propose a stationarity test based on multiscale, locally stationary, 2D wavelets. Three separate experiments were performed to evaluate the capabilities and the limitations of this test. The experiments comprised a chessboard stationarity analysis, two classification tasks, and a psychophysical experiment. The classification tasks were performed on 110 texture images from a texture database. In one subtask, five texture feature vectors were extracted from each image and the classification accuracy of two classical methods compared, whereas in the second subtask, the classification accuracy of several methods was compared to the descriptors defined for each image within the database. In the psychophysical experiment, the correlation between the classification results and observer judgements of texture similarity were determined. It was found that a combination of wavelet shrinkage and rotation-invariant local binary pattern best predicted the observer response. The results show that the proposed stationarity test is able to provide relevant information for texture analysis. © *2021 SPIE and IS&T* [DOI: 10.1117/1.JEI.30.4.043001]

Keywords: texture stationarity; visual texture; texture classification.

Paper 200833 received Dec. 5, 2020; accepted for publication Jun. 15, 2021; published online Jul. 2, 2021.

1 Introduction

A one-dimensional temporal signal is said to be stationary if its local statistical properties are constant in time.[1] The same concept can be extended to the field of texture analysis, where it is widely used. It is, in fact, a fundamental assumption of global texture models, such as Markov random fields, autocorrelation functions, and the well-known Tamura features.[2] Consequently, if the texture is not stationary, these techniques give a flawed representation of the signal. In fact, this fails to depict the actual mathematical properties of the texture, because it extracts an average behavior that neglects any change of local features in the image. For such a case, a texture can be divided into stationary subregions with a segmentation algorithm that autonomously partitions an image into multiple homogeneous areas.[2] However, this increases the computational complexity of the analysis and, as discussed in Ref. 3, prior knowledge of the stationarity of the sample would still be needed.

Currently, the stationarity property of a texture image has a dual meaning. From a mathematical point of view, the term "stationarity" means that the average of the data generating process, which gives rise to the image observed, is the same everywhere in the image,[3] and that its distribution is essentially regular, i.e., its variance is finite and its covariance is dependent only on the distance between pixels. We use the term "data stationarity" to refer to this definition. On the other hand (in Ref. 4, pg. 80), Petrou and Sevilla stated that "a stationary texture image is an image which contains a single type of texture." This suggests an interpretation of stationarity that is more related to the human visual system. We name this second definition "visual stationarity." This interpretation is more complex than the first one, as it touches the border of linguistics, i.e., the understanding of the concept of "a single type of texture," which would probably depend on the context of analysis. This situation is not uncommon, for example, even the similar but more widely used term "homogeneous" has a fuzzy meaning in the research of human perception of texture.[5]

*Address all correspondence to Michele Conni, michele.conni@barbierielectronic.com

1017-9909/2021/$28.00© 2021 SPIE and IS&T

An example that shows the practical need for a shared definition of data and visual stationarity is suggested by Bello-Cerezo et al.[6] In this paper, the authors classified a wide variety of existing texture databases according to certain characteristics. Among these characteristics is texture stationarity, with explicit mention to the definition given in Ref. 3. In Appendix C, we show that this partition is inconsistent with the results of the stationarity test employed in this paper (see Sec. 2). This mismatch clearly highlights the discrepancy between data and visual stationarity.

In a previous study,[7] we analyzed the concept of data stationarity and we expanded a framework developed to evaluate image stationarity[8] to account for multiple scales. The choice of scale has proven to be of great importance for texture analysis.[9] The link between this mathematical approach and visual stationarity is discussed in Ref. 8, in which the stationarity test was applied to images of pilled fabric, and the results subsequently compared to the authors' evaluation of the visual stationarity of the images. This, however, provides limited experimental psychophysical data on which to base firm conclusions. In this work, we wish to remedy this gap in the literature. To this end, we provide an investigation of the relationship between visual and data stationarity, using images of our own dataset as well as from a subset of the describable texture dataset (DTD).[10] The choice of DTD is based on its texture categorization and annotation by human observers, thus incorporating the complexity of human perception of texture. Additionally, data stationarity analyses of an alternative texture database can be found in Appendix C.

2 Texture Stationarity

In the field of visual texture analysis, the conjecture proposed by Julesz, stating that "whereas textures that differ in their first- and second-order statistics can be discriminated from each other, those that differ in their third- or higher-order statistics usually cannot,"[11] is a good approximation of how human perception works. Many texture feature extraction techniques, therefore, assume that their image targets are second-order stationary,[3,12] i.e., the process generating these images has a constant mean, a finite variance, and a covariance that is a function of pixel distance $\mathrm{cov}(X_{\mathbf{r}_1}, X_{\mathbf{r}_2}) = \gamma(\mathbf{r}_1 - \mathbf{r}_2)$. Taylor et al.[8] employed these premises to develop an image stationarity test for a single realization of a generating statistical process. The test interprets each image as a locally stationary two-dimensional wavelet (LS2W) process, and it evaluates the constancy of its power spectrum to estimate its stationarity. We introduce and describe it in Sec. 2.1, and we propose a variation to it in Sec. 2.2. Frequently used mathematical notations are also summarized in Table 1.

2.1 *LS2W Stationarity Test*

A mother wavelet $\psi(x)$ is a compact support function with oscillatory characteristics,[13–15] $x \in \mathbb{R}$, which, together with an auxiliary function $\phi(x)$ called a father wavelet, can be used to form a complete functional basis on $L^2(\mathbb{R})$. This functional basis $\{\psi_{j,k}, \phi_{j,k}\}_{j,k\in\mathbb{Z}}$ is achieved by scaling and shifting $\psi(x)$ and $\phi(x)$, with j and k indicating the scaling and shifting indices, respectively. On the one hand, the shifting gives the possibility of representing local segments of the signal, whereas on the other hand, the scaling allows it to represent the fine or coarse structures contained therein. As discussed in Ref. 16, a discrete version of such a basis function can be obtained by associating two compactly supported mother and father wavelets ψ and ϕ with a suitable pair of low-pass and high-pass filters $\{h_k\}_{k\in\mathbb{Z}}$ and $\{g_k\}_{k\in\mathbb{Z}}$. In this case, a discrete wavelet at scale j is a vector $\psi_j = (\psi_{j,0}, \dots, \psi_{j,N_j-1})$, where $N_j = (2^j - 1)(N_h - 1) + 1$, $N_h \neq 0$, $\psi_{-1,n} = g_n$, and $\psi_{j-1,n} = \sum_k h_{n-2k}\psi_{j,k}\ \forall\ n \in [0, \dots, N_{j-1} - 1]$. Such a basis can be easily expanded in two dimensions and applied to images. This is achieved by substituting k with $\mathbf{k} = (k_1, k_2)$ and introducing a direction index $l \in \{H, V, D\}$. l is employed to mix both father and mother wavelets, to ensure the completeness of the basis. Its values are H for horizontal, V for vertical, and D for diagonal. The corresponding 2D fundamental wavelets are defined as $\psi^H_{j,\mathbf{k}} = \phi_{j,k_1}\psi_{j,k_2}$, $\psi^V_{j,\mathbf{k}} = \psi_{j,k_1}\phi_{j,k_2}$, and $\psi^D_{j,\mathbf{k}} = \psi_{j,k_1}\psi_{j,k_2}$. A generic discrete wavelet at scale j in a given decomposition direction l can then be expressed as in Eq. (1):

Table 1 Frequently used mathematical notations.

$\psi(x)$, $\phi(x)$	Mother and father wavelets, respectively
γ	Covariance function $\gamma: \mathbb{Z}^2 \rightarrow \mathbb{R}$ of a stationary image
$\mathbf{r}_i$	Location or coordinate of arbitrary pixel i, with $\mathbf{r} \in \mathbb{Z}^2$
j	Wavelet scaling index, where $j \in \mathbb{Z}^+$
k, $\mathbf{k}$	Wavelet shifting index and vector, respectively, $k \in \mathbb{Z}$ and $\mathbf{k} = (k_1, k_2)$, $k_i \in \mathbb{Z}$
$\mathbf{u}$	Image coordinates after wavelet shifting, $\mathbf{u} = \mathbf{r} + \mathbf{k}$
l	Wavelet direction index, where $l \in \{H, V, D\}$
h, g	Discrete low- and high-pass filters, respectively
N_h	Number of non-zero elements in h, $N_h = \#(h) \neq 0$
N_j	Number of elements of the discrete wavelet ψ_j at scale j
R, C	Number of rows and columns in an image, respectively, expressed in terms of a power of 2, $R = 2^m$, $C = 2^n$, n, $m \in \mathbb{N}^+$
$\mathbf{R}$	Dimension of a grayscale image, $\mathbf{R} = (R, C)$
$w^l_{j,\mathbf{u}}$	Coefficient of the wavelet transform
$\xi^l_{j,\mathbf{u}}$	Zero-mean random orthonormal increment sequence
J	Lowest significant scale, $J(R, C) = \log_2\{\min(R, C)\}$
$X_{\mathbf{r};\mathbf{R}}$	Generic LS2W process with dimension $\mathbf{R}$
$\mathbf{z}$	Normalized spatial coordinate, $\mathbf{z} = \mathbf{u}/\mathbf{R} := (u/R, v/C)$, $\mathbf{z} \in (0,1)^2$
S^l_j	Local wavelet spectrum
$d^l_{j,\mathbf{u}}$	Empirical mother wavelet coefficients
$\mathbf{I}(\mathbf{u})$	LWP as an estimator for S^l_j
A_J	LWP correction matrix
$\hat{\mathbf{S}}(\mathbf{u})$	Estimator for LWP, $\hat{\mathbf{S}}(\mathbf{u}) = A_J^{-1}\mathbf{I}(\mathbf{u})$, composed by the elements $\hat{S}^l_j(\mathbf{u})$
T_{ave}	Departure from constancy of an estimated LWP $\hat{\mathbf{S}}(\mathbf{u})$
B	Number of repetitions of the bootstrap loop
p	p value of the stationarity test for $\hat{\mathbf{S}}$
$\eta(j, l)$	Index of scale-direction pair, $\eta \in \{1, \dots, 3J\}$
$p_{\eta(j,l)}$	p value of the stationarity test for $\hat{S}^l_j$
$\mathbf{p}$	Vector of $p_{\eta(j,l)}$ at various dyadic scales and directions, $\mathbf{p} = (p_{\eta=1}, \dots, p_{\eta=3\,J})$
p_j	p value of the stationarity test for $\hat{S}_j$
$\mathbf{p}_j$	Vector of p_j at various dyadic scales, $\mathbf{p}_j = (p_{j=1}, \dots, p_{j=J})$
$X_{wn}(\mathbf{r})$	Bidimensional white-noise process
$N(\mu, \sigma)$	Normal distribution with mean μ and standard deviation σ

$$\psi_j^l = \begin{bmatrix} \psi_{j,(0,0)}^l & \cdots & \psi_{j,(0,N_j-1)}^l \\ \vdots & \ddots & \vdots \\ \psi_{j,(N_j-1,0)}^l & \cdots & \psi_{j,(N_j-1,N_j-1)}^l \end{bmatrix}. \tag{1}$$

The family of wavelets $\{\psi_j^l\}$ derived from the definition of Eq. (1) was used in Ref. 17 to define a random field modeling framework, called the LS2W field. The idea is to apply the complete set of 2D discrete wavelet matrices as filters on an image to calculate its wavelet coefficients in various pixel positions, determined by the shift $\mathbf{k}$. The approach proposed in Ref. 17 exploits only dyadic scales, whereas the filters are applied on every possible position of the image. Mathematically, a generic image of dimensions $\mathbf{R} = (R, C)$ can be generated with an LS2W process as in Eq. (2), where $\{w_{j,\mathbf{u}}^l\}$ are the wavelet coefficients and $\{\psi_{j,\mathbf{u}}^l(\mathbf{r}) = \psi_{j,\mathbf{u}-\mathbf{r}}^l\}$ are 2D discrete non-decimated wavelets with orientation l, scale j, and shifted coordinate $\mathbf{u}$. Each coefficient $w_{j,\mathbf{u}}^l$ quantifies how large the contribution of the corresponding wavelet $\psi_{j,\mathbf{u}}^l(\mathbf{r})$ is in defining the process. $\{\xi_{j,\mathbf{u}}^l\}$ is a zero-mean random orthonormal increment sequence, which allows stochastic structure to be encapsulated in the model. The dependence on the image dimension $\mathbf{R}$ is included to make the link with the lowest significant scale $J(R, C) = \log_2\{\min(R, C)\}$ explicit. Further on, it will be considered as implicit.

$$X_{\mathbf{r};\mathbf{R}} = \sum_l \sum_{j=1}^{\infty} \sum_{\mathbf{u}} w_{j,\mathbf{u};\mathbf{R}}^l \psi_{j,\mathbf{u}}^l(\mathbf{r}) \xi_{j,\mathbf{u}}^l. \tag{2}$$

The local wavelet spectrum (LWS) of an LS2W process $X_\mathbf{r}$ can be considered as a power spectral density for the stationary wavelet transform $S_j^l(\mathbf{z}) \approx w_j^l(\mathbf{u}/\mathbf{R})^2$. Here $\mathbf{z} \in (0,1)^2$ is a normalized spatial coordinate $\mathbf{z} = \mathbf{u}/\mathbf{R} \coloneqq (u/R, v/C)$ and, for a stationary process, $S_j^l(\mathbf{z})$ is a constant function of $\mathbf{z}\ \forall\ j, l$.[17] Therefore, an estimate of the LWS can be used to assess the stationarity of an image.[8]

Eckley et al.[17] proposed the local wavelet periodogram (LWP) as an estimator for $S_j^l(\mathbf{z})$. It is expressed as in Eq. (3), where $d_{j,\mathbf{u}}^l$ are the empirical mother wavelet coefficients of the image. This fact that the father wavelet coefficients are not included in Eq. (3) implies the independence of the LWP from the mean value of the process under study. This estimator is biased, but it can be corrected by multiplying it with the inverse of the two-dimensional discrete autocorrelation wavelet matrix A_J, obtaining $\hat{\mathbf{S}}(\mathbf{u}) = A_J^{-1}\mathbf{I}(\mathbf{u})$.[18] Prior to the correction, wavelet shrinkage has also been applied to each level of LWP to increase the consistency of the estimator.[17] $\hat{\mathbf{S}}(\mathbf{u})$ is an array with four dimensions, i.e., two for the spatial coordinates $\mathbf{u}$, one for scale j, and one for direction l, to reach a total of $R \times C \times J \times 3$ elements.

$$\mathbf{I}(\mathbf{u}) = \{I_{j,\mathbf{u}}^l\} = \{|d_{j,\mathbf{u}}^l|^2\} = \left\{\left(\sum\nolimits_{\mathbf{r}} X_\mathbf{r} \psi_{j,\mathbf{u}}^l(\mathbf{r})\right)^2\right\}. \tag{3}$$

The stationarity test introduced in Ref. 8 employs as test statistic a departure from constancy $T_{\text{ave}}\{\hat{\mathbf{S}}\}$ in Eq. (4), which is the variance of the values of $\hat{\mathbf{S}}(\mathbf{u})$, averaged over scales j and directions l:

$$T_{\text{ave}}\{\hat{\mathbf{S}}\} = (3J)^{-1} \sum_l \sum_{j=1}^{J} \text{var}_\mathbf{u}(\hat{\mathbf{S}}(\mathbf{u})). \tag{4}$$

Since the original distribution is unknown *a priori*, i.e., the algorithm operates on a single realization of the LS2W process $X_\mathbf{r}$, it is simulated with a bootstrap operation to infer its characteristics from the input image. Then the p value of the stationarity test is calculated by comparing the T_{ave} of the various bootstrap iterations with that of the original image. Mathematically, this can be expressed as $p = \frac{1+\#\{T_{\text{ave}}^{\text{obs}} \le T_{\text{ave}}^{(i)}\}}{B+1}$, where obs is indicating the observed image, index i is specifying the various bootstrap instantiations, and B is the total number of repetitions of the bootstrap loop.

2.2 *Proposed Approach to Texture Stationarity*

As discussed in Ref. 8, it is also possible to test each scale-direction spectral plane for constancy. This is achieved by defining a test statistic $T_{\eta(j,l)}$ as shown in the following equation:

$$T_{\eta(j,l)}\{\hat{\mathbf{S}}\} \coloneqq T\{\hat{S}_j^l\} = \mathrm{var}_{\mathbf{u}}(\hat{S}_j^l(\mathbf{u})). \tag{5}$$

With these test statistics, it is possible to perform a stationarity test at every scale j and direction l. For each of these tests, a p value $p_{\eta(j,l)} = \frac{1+\#\{T^{\mathrm{obs}}_{\eta(j,l)} \leq T^{(i)}_{\eta(j,l)}\}}{B+1}$ can be defined. The $p_{\eta(j,l)}$'s can be grouped into a vector $\mathbf{p}$ to understand the degree of stationarity of an image at dyadic scales $2^j \ \forall\ j \in \mathbb{Z}^+, j < J(R, C)$ and for direction $l \in \{H, V, D\}$. Note that the calculation of p is nonlinear in T_{ave}, which means that the average value of the vector $\mathbf{p}$ is different from the p value of the image ($\overline{p_{\eta(j,l)}} \neq p$). Given that the family of wavelets $\{\psi_j^l\}$ is composed by orthogonal filters, each test is independent from the others. Such an approach is similar to the $\texttt{Bootstat}^{\texttt{mh}}_{\texttt{LS2W}}$ framework introduced in Ref. 8, which, however, is used to probe the stationarity of the whole image and not scale by scale. To achieve that, the $\texttt{Bootstat}^{\texttt{mh}}_{\texttt{LS2W}}$ applies a multiple hypothesis testing scenario discussed in Ref. 19. In our case, it is not necessary to resort to this correction method since we define $3J$ distinct hypothesis tests, one for each η.

Finally, we can define a set of p values under the null hypothesis of stationarity of scale j. We can define these as $p_j = \frac{1+\#\{T_j^{\mathrm{obs}} \leq T_j^{(i)}\}}{B+1}$, $T_j\{\hat{\mathbf{S}}\} \coloneqq \frac{T_{\eta(j,H)}\{\hat{\mathbf{S}}\}+T_{\eta(j,V)}\{\hat{\mathbf{S}}\}+T_{\eta(j,D)}\{\hat{\mathbf{S}}\}}{3}$. This corresponds to Eq. (8) of Ref. 8, averaged only over the wavelet directions. We gather these values in the vector $\mathbf{p}_j$.

p, $\mathbf{p}$ and $\mathbf{p}_j$ are all results of statistical tests, under the null hypothesis of stationarity. Therefore, a threshold α of the test significance level can be chosen. In this paper, we used Haar wavelets and we set $B = 100$ and $\alpha = 5\%$, where not stated otherwise.

3 Chessboard Stationarity

The relevance of LS2W and the related stationarity test in computer vision applications has been discussed in various papers.[8,17,20,21] However, to our knowledge, its relationship to the human perception process and appearance analysis has not been previously addressed. A simple way to obtain some initial insights is by extracting the values p and $\mathbf{p}$ from images that have a regular structure. Our goal is, therefore, to investigate whether the tests introduced in the previous section define these images as data stationary.

We applied the stationarity test on an 8-bit 128×128 chessboard image [Fig. 1(a)] with binary values (0 on black patches and 255 on white ones). Following the classification discussed in Ref. 3, this is a regular periodic pattern with a black square primitive, which thus can be analyzed with a shape grammar. According to the definition in chapter four of the same reference, the image can be perceived as filled with a single texture, i.e., it is visually stationary. Since the LS2W model assumes that the image analyzed has a stochastic structure, we added a degree of random noise to the pixels of the chessboard picture: a two-dimensional white Gaussian noise process $X_{wn}(\mathbf{r}) \sim N(0, \sigma_{wn})$, where $\sigma_{wn} = 10 \cdot n \ \forall\ n \in [0{,}10]$. Only moments of order two and higher of the original checkerboard picture are influenced by the procedure because the distribution has an average value of zero. Given that for all of these images the p value of the LS2W test is $p = 1$, they are data stationary according to the test. In this case, the scale analysis does not add any additional information, because $\mathbf{p}$ is constant and unitary $\mathbf{p} = \{p_{\eta(j,l)} = 1 \ \forall\ j \in \mathbb{Z}^+ \ \forall\ l \in \{H, V, D\}\}$. Interestingly, the results of the stationarity tests are the same for both the stochastic ($\sigma_{wn} \neq 0$) and the deterministic ($\sigma_{wn} = 0$) patterns that were analyzed. This is probably due to the fact that the added noise is second-order stationary and thus does not influence the result of the calculation, which is mainly dictated by the deterministic base. The zero mean of the overall image, which is assumed by the LS2W methodology, is, as discussed, independent on the noise distribution. This property is further ensured by the definition of the LWP itself [see Eq. (3)], which is an estimator of the LWS of an image. In fact, as mentioned in Sec. 2.1, the LWP neglects the constant component of the image contained in the

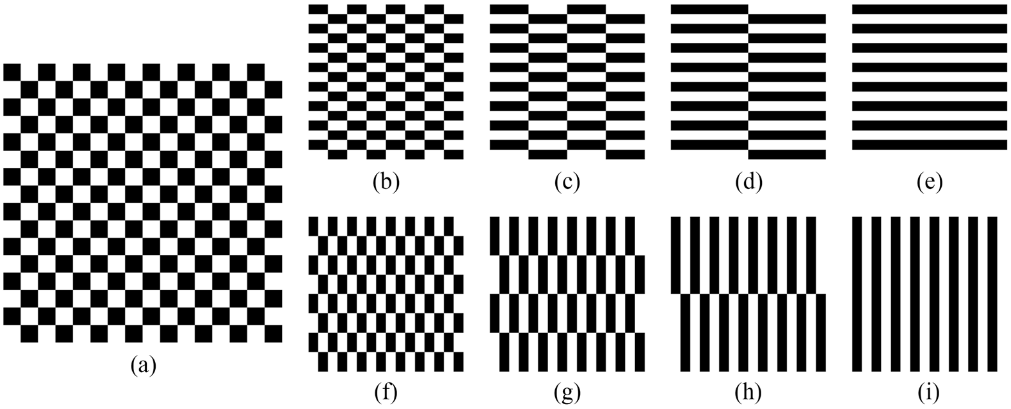


Fig. 1 (a) Original chessboard image and (b)–(e) horizontally and (f)–(i) vertically stretched versions.

father wavelet coefficients. Based on these results, we avoid adding a stochastic structure to deterministic images in the following analysis.

The stationarity analysis has also been repeated on modified versions of the chessboard image in order to understand how various types of distortions affect **p**. First, the image is stretched in the horizontal and vertical directions [see Figs. 1(b)–1(i)]. Their p values appear to be always 1 and thus unaffected by the stretching. The corresponding **p** is also mainly unitary. The only exceptions are the repeatable drops at scales 2^1 [Fig. 1(c)] and 2^2 [Fig. 1(g)] shown in Fig. 2. These effects arise from $T_{\text{ave}}\{\hat{\mathbf{S}}\}$ having a peak in the diagonal direction, which is stronger than in any other image. This peak is present in every image, at scale 2^2 in the horizontal and vertical directions and 2^3 in the diagonal one. However, in the case of Figs. 1(c) and 1(g), it is particularly strong in respect with the values at other scales. In fact, in these levels and with these images, the trade-off between spatial distance and frequency of the changes in intensity is the highest. Given that the stretched images appear visually stationary, this unexpected effect highlights a limitation of the mathematical method.

Next, we rotated the chessboard image to confirm the independence of the test from the direction of the texture. This should be ensured by the completeness of the family of filters $\{\psi_j^l\}$ considered, as they probe all the relationships between pixels at the scale j. We used the function `imutils.rotate` from Python,[22] based on bilinear interpolation, with rotation angles [0 deg, 90 deg), in an interval of 10 deg. Bigger angles of rotation are not necessary, given that the original chessboard image is symmetric by rotation of π and the horizontal and

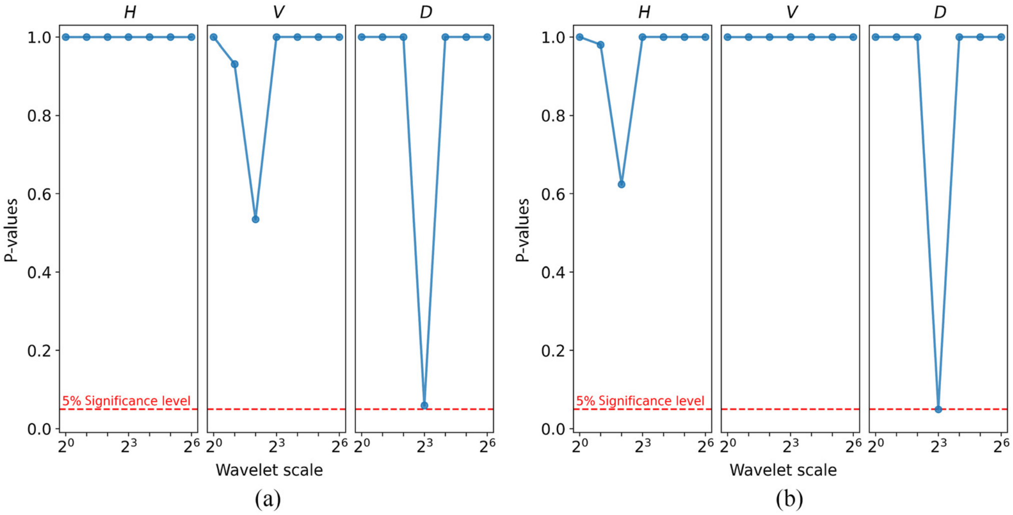


Fig. 2 (a) **p**'s for Fig. 1(c) and (b) Fig. 1(g) (H for horizontal, V for vertical, and D for diagonal).

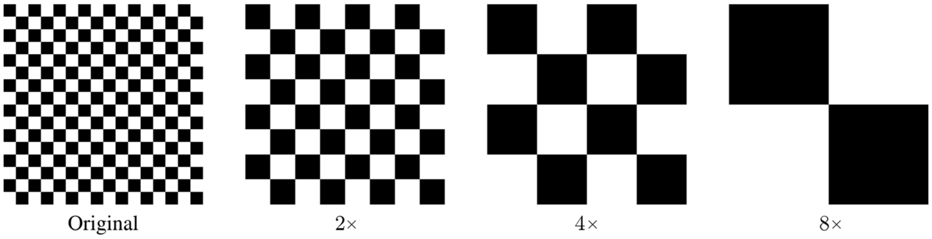


Fig. 3 Original image and resized variants.

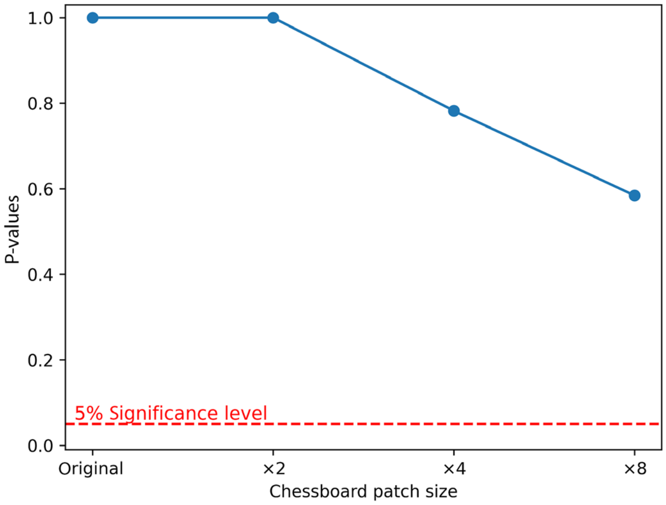


Fig. 4 Dependence of chessboard image p value on the size of its patches.

vertical wavelets probe perpendicular directions. As expected, all these images appeared to be data stationary at every scale considered for each rotation angle.

Finally, we varied the sizes of the white and black patches in the chessboard. To maintain the original image dimension (128×128), the sides of the patches were enlarged by powers of 2, as shown in Fig. 3. Again, all the images appear to be stationary with $\alpha = 5\%$. However, it is interesting to notice that the p is not unitary for all the images, as shown in Fig. 4. Visually, this can be linked to a reduction of homogeneity of the whole image associated with the scaling. In parallel, an analysis of the vector $\mathbf{p}$ of the images, shown in Fig. 5, displays an average decrease of values with the zoom, especially in the finer scales. This could be due to the different cone of influence (COI) of the wavelet at each scale: the finer wavelets, whose COIs are the smallest

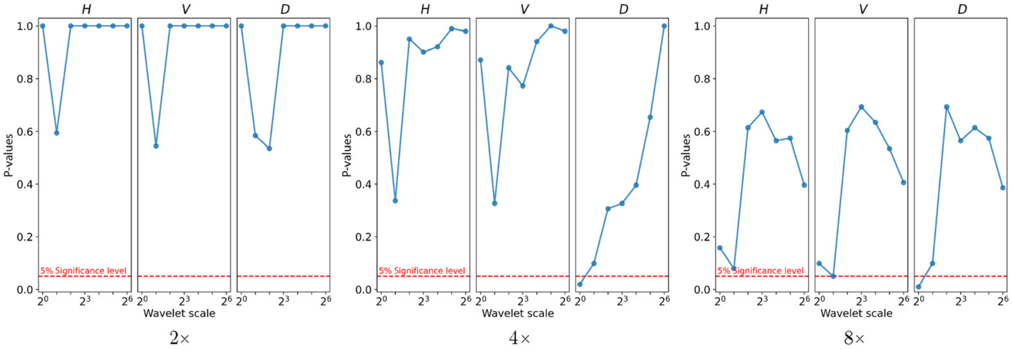


Fig. 5 Vectors $\mathbf{p}$ for 2×, 4×, and 8× resized chessboards showed in Fig. 3. Note that the y axes are identical and the labels over each graph indicate the direction of the wavelet (H for horizontal, V for vertical, and D for diagonal).

considered, are affected by the edges of the pattern more abruptly than the other scales, which could lead to a smaller stationarity.

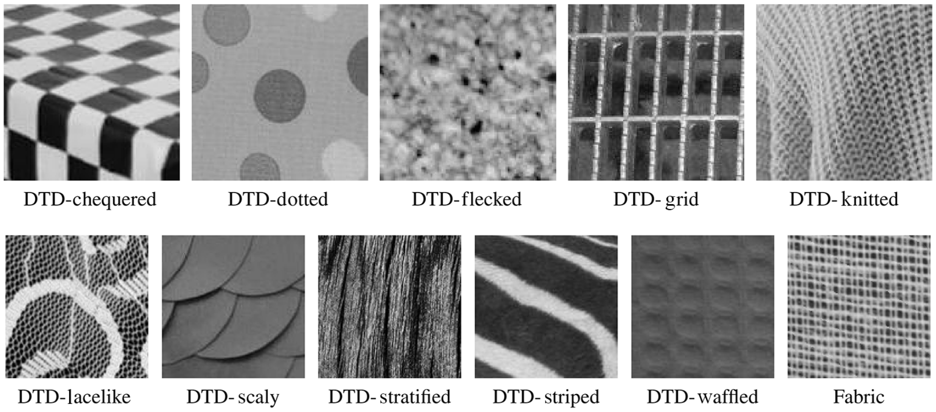
4 Texture Classification

To further probe how the stationarity information can be linked to the perception of texture, we ran a classification experiment. We used the DTD,[10,23] whose purpose is to describe textures "in the wild" with semantic attributes chosen by human observers. We selected 100 texture images among the ones in the database, extracting them from 10 different classes derived from Ref. 5. These have been then supplemented with an eleventh class, consisting of 10 pictures of fabric samples, which we acquired ourselves, for a total of 110 images. The limited number of images considered is bounded by the necessity of submitting them to the observers of the psychophysical experiment discussed in Sec. 5. The classes are: chequered, dotted, fabric, flecked, grid, knitted, lacelike, scaly, stratified, striped, and waffled. Although the original images are in color, in this experiment, they have been converted to greyscale images, in order to account only for their spatial variation. This conversion has been performed by loading each image with the `cv2.imread` function of the OpenCV library, which derives the intensity information Y as $Y = 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B$.[24] Every picture has also been cropped into squares of 128×128 pixels, as required by the current implementation of the algorithm. Examples of the selected and processed images from all 11 classes are shown in Fig. 6.

We ran two classification tasks on these data. For the first task, we divided each image input into subimages. To comply with the requirements of the dyadic implementation of LS2W, each image target was split into 16 subimages, each of dimension 32×32 pixels (for additional insight on the choice of dimension, see Appendix D). Then we classify these subimages and evaluate whether according to the algorithm they belong to their original image. Details on the experiment setup for the first task can be seen in Fig. 7. As for the second task, all the 110 images were classified using more varied texture features and the previously mentioned DTD groups as ground truth classes. Details for this experiment are given in Sec. 4.3.

4.1 *Texture Feature Extraction*

Before presenting the results of the classification tasks, we introduce the texture features used for them. Five texture feature vectors are extracted from each image, each considered at the seven dyadic scales (five for the subimages) used to extract its corresponding **p**. Feature extraction methods have been selected from the collection of techniques considered in Ref. 6. We neglected the non-scalable approaches and the learning-based ones. We did not consider the latter because, without additional training, the off-the-shelf learning models are also non-scalable, and also because they have not been discussed by Ref. 3 when defining stationary textures.


DTD-chequered DTD-dotted DTD-flecked DTD- grid DTD- knitted

DTD-lacelike DTD- scaly DTD- stratified DTD- striped DTD- waffled Fabric

Fig. 6 Examples of images used in the experiment, coming from all 11 classes or categories of images. 10 classes originate from DTD[10] and one comes from our own dataset of white fabrics.

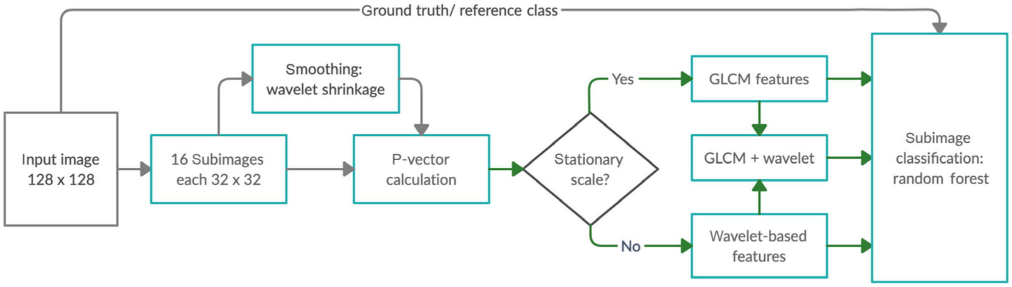

Ground truth/ reference class
Input image 128 x 128
16 Subimages each 32 x 32
Smoothing: wavelet shrinkage
P-vector calculation
Stationary scale?
Yes
No
GLCM features
GLCM + wavelet
Wavelet-based features
Subimage classification: random forest

Fig. 7 Experimental setup for the mixed subimage classification task in Sec. 4.2. In this, each input image is split into subimages, and the latter is classified and evaluated as whether they belong to their original or source image.

However, we tested as reference the capabilities of off-the-shelf models in Sec. 4.3. Details of the feature extraction approaches used and their parameters are as follows.

Rotation-invariant local binary pattern (LBP)[25] vectors are obtained at dyadic radii of 2^n, $n \in [0,7]$ and eight angles $(k \cdot \pi/4)$ with interpolation. For each scale and image, an LBP histogram of 36 elements is then used as its feature vector.

Gray-level co-occurrences matrix (GLCM)[26] vectors are obtained. They are composed of five global statistical parameters (contrast, correlation, energy, entropy, and homogeneity) at four angles (0, $\pi/4$, $\pi/2$, and $3/4\pi$). The feature vector is, therefore, 20 elements long for each scale.

Histogram of oriented gradients (HOG)[27] is computed at nine orientations, normalized according to the hysteresis L2-norm, with 2^n pixels per cell and one cell per block. The HOG vectors thus obtained have a total of nine elements per scale.

Gabor filters-based features[28] with central filter frequencies $f = 1/2^n$, $n \in [0,7]$, eight orientations, and deviation parameters γ and η assumed to be equal to $3\,\ln(2)/2\pi$. γ and η are chosen such that half-peak magnitude iso-ellipses of the various filters would not overlap (see Appendix C of Ref. 29 for more details). Input images are filtered with these and their mean and variance has been calculated, resulting in a 16-element vector for each scale.

Wavelet vectors generated with Haar and Daubechies filters at dyadic scales. These features correspond to methods I and III used in Ref. 17. As in the Gabor-based ones above, the mean and variance energy of the filtered images have been used. Note that the variance of the energy has some degree of correlation with the LWP [Eq. (3)] and with the test statistics adopted [Eqs. (4) and (5)]. However, the two have some substantial differences: while the regular wavelets used to extract the features are placed at dyadic locations in the image, the LS2W model used by the test is based on non-decimated discrete wavelets. Moreover, the estimator $\hat{\mathbf{S}}(\mathbf{u})$ is corrected with the discrete autocorrelation wavelet matrix A_J. Horizontal, vertical, and diagonal wavelets have been considered, such that each feature vector has a length of 12.

According to Refs. 3 and 6, within this list, GLCM and HOG are texture features better suited for stationary texture images, whereas the others are better at characterizing non-stationary ones. This suggests that it could be possible to evaluate the stationarity of an image based on the other features. This, however, could be achieved only by defining a proper testing procedure for each methodology.

4.2 *Classification Task 1: Mixed Subimage Classification*

The vector $\mathbf{p}$ by itself reflects the stationarity (or lack of it) of an image. Therefore, it is not able to wholly represent the peculiar characteristics of a texture by itself, which is what is required by the features used for classification. Nonetheless, it is possible to use the stationarity information to optimize the process of texture feature extraction. In fact, some feature extraction techniques are claimed to be more appropriate for non-stationary images than others,[3,6] although such a claim has not been proven experimentally. According to this idea, the stationarity information contained in the $\mathbf{p}$ could suggest a selection of features at different scales, which is optimal to describe the texture. In the context of classification, this translates into an increase in the accuracy of the process. In this section, we set out to provide experimental proof for this hypothesis.

The experiment setup for this classification task can be seen in Fig. 7. The choice of features for stationary and non-stationary texture is based on Ref. 3. Both GLCM and wavelet-based approaches were used to extract features at dyadic scales, allowing us to classify the subimages using each individual feature vector. We also combine GLCM and wavelet-based features based on the $\mathbf{p}$ of each image. One of the methods, indicated with f_s, is applied to stationary scales and the other, referred to with f_{ns}, to non-stationary ones. The j'th element of the mixed feature vector f_{mix} is then obtained as

$$f_{\text{mix},j}(f_s, f_{ns}) = \begin{cases} \text{pad}(f_{s,j}) & \text{if } p_j > \alpha \\ \text{pad}(f_{ns,j}) & \text{if } p_j < \alpha \end{cases}, \tag{6}$$

where $f_s \neq f_{ns}$ and $p_j \in \mathbf{p}_j$. The hypothesis behind the calculation of f_{mix} is that the only features affected by the non-stationarity of a certain scale would be the ones at that same scale. In the present case, $f_s = \text{GLCM}$ and $f_{ns} = \text{wavelet}$. The threshold α, which in the current work is set to 5%, is applied to the values p_j to estimate whether each scale is stationary. If the p value at a certain scale is bigger than α, the image is considered stationary at that scale and the GLCM feature vector is inserted in the mixed vector. Otherwise, the image at that scale is considered non-stationary and the wavelet-based feature is used. In this way, the space of the mixed vector can be divided into a stationary subspace and a non-stationary one, each one orthogonal to the other. As a note, $f_{\text{mix},j}$ is padded with zeros to the right in the stationary case and to the left in the non-stationary one so that the length of the mixed vector is equal to the sum of the lengths of the other two vectors.

As shown in Fig. 7, we chose to use a random forest classifier with a 67% to 33% training-test set subdivision. The forest has 100 trees and the algorithm selects their depth so that the nodes are expanded until all the leaves are pure. At each split of the tree, the square root of the initial number of features is considered. For every process, the classification has been repeated 1000 times and the average value of the accuracy was extracted. The results of using GLCM and wavelet-based features individually as well as in a mixed feature vector are shown in Table 2. Combining the two techniques appears to worsen the classification accuracy. The use of wavelet as f_s and GLCM as f_{ns} is discussed in Appendix B. Appendix B also shows how the use of wavelet shrinkage in the extraction of the $\mathbf{p}$, discussed in Sec. 2.1, leads the accuracy to drop to 61.81%. Additional analyses of alternative classification experiments are also reported in Appendices A and E.

4.3 *Classification Task 2: Mixed DTD Classification*

As a second task, we classified the unabridged images on the basis of the classes defined by the DTD authors.[5] In this case, we used all the features extraction techniques described in Sec. 4.1, so to probe a wider range of possible approaches. As in the previous section, we derived the accuracy obtained both by classifying the dataset with the original features and with all the possible combinations of mixed vectors f_{mix} [see Eq. (6)]. The results for the classification without shrinkage are reported in Table 3, where, according to the practice adopted in Eq. (6), f_s indicates the feature extraction method considered as stationary, whereas f_{ns} indicates the non-stationary one. On the diagonal of the tables, we show as reference the results for the

Table 2 Classification accuracy corresponding to the classification task 1 shown in Fig. 7.

Method	GLCM	Wavelet	f_{mix}
Average accuracy (in %)	68.05	65.33	64.62

Table 3 Classification accuracy corresponding to the classification task 2 without wavelet shrinkage.

		f_{ns}				
		LBP (%)	GLCM (%)	HOG (%)	Gabor (%)	Wavelet (%)
f_s	LBP	32.7	32.4	31.6	32.8	32.3
	GLCM	32.5	31.7	31.3	31.4	31.9
	HOG	27.5	26.2	23.4	25.9	25.9
	Gabor	27.4	27.3	27.6	26.2	27.3
	Wavelet	33.8	32.0	31.4	32.5	29.9

Table 4 Classification accuracy corresponding to the classification task 2 with wavelet shrinkage.

		f_{ns}				
		LBP (%)	GLCM (%)	HOG (%)	Gabor (%)	Wavelet (%)
f_s	LBP	32.7	34.0	30.3	30.2	33.6
	GLCM	33.8	31.7	31.6	29.2	33.3
	HOG	33.7	33.2	23.4	26.4	33.1
	Gabor	34.7	32.7	33.9	26.2	32.4
	Wavelet	36.6	34.4	31.1	33.9	29.9

classification without mixing. The results of applying shrinkage to the same set of experiments are also shown in Table 4.

These results are compatible with those obtained by local descriptors for the whole DTD.[10] Between the unmixed features of Tables 3 and 4, the LBPs are the most successful. On the other hand, HOG and Gabor features appear to perform quite poorly. The mix that provides the best classification accuracy is $f_s = \text{wavelet}$ and $f_{ns} = \text{LBP}$. In general, the mixing appears to improve the performance of the classification with every technique. In this case, applying the wavelet shrinkage when calculating $\mathbf{p}_j$ seems to be the best choice.

The accuracy depends more on the stationary technique (f_s) than on the non-stationary one (f_{ns}). This is due to the fact that the images chosen are mainly stationary: the 88% of the p_j's are bigger than the 5% test threshold without wavelet shrinkage, whereas the percentage drops to 77% with wavelet shrinkage. The whole-image p values have a similar statistic, with 90% for the rough and 92% for the smooth. This could be related to the fact that the test used is conservative.[8]

As an additional reference, we performed the same experiment with the following seven convolutional neural networks (CNNs):

- ResNet-50,[30]
- VGG-16 and VGG-19,[31]
- Inception v3,[32]
- DenseNet-121, DenseNet-161, and DenseNet-201.[33]

We extracted the features from off-the-shelf models, which were trained for object recognition. Each network was employed as a generic feature extractor, and the resulting features were then passed on to the random forest classifier. Every individual network extracts 1000 features per image. Given that the CNNs required the input array to have certain dimensions and to have three channels, we resized them accordingly using cubic interpolation and we tripled

Table 5 Classification accuracy with CNNs corresponding to the classification task 2.

Network	Average accuracy (%)
ResNet-50	44.0
VGG-16	39.7
VGG-19	47.4
Inception v3	40.5
DenseNet-121	43.3
DenseNet-169	42.0
DenseNet-201	42.7

the gray channel. The results are shown in Table 5, where one can see that this assignment is challenging even for learning-based techniques.

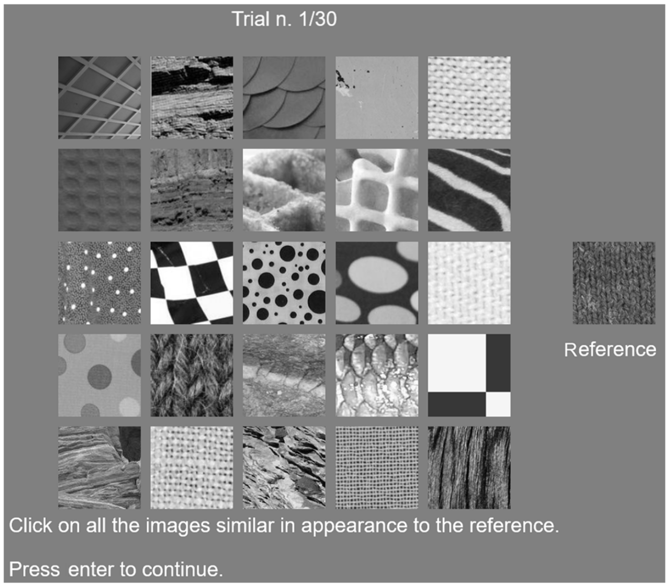
5 Psychovisual Experiment Design

As final step of our investigation of the link between perceptual and data stationarity, we performed a psychophysical experiment. We used the psychovisual software PsychoPy2[34] to set up the experiment and uploaded it to the Pavlovia web platform.[35] The images used in the experiment are the same 110, which were classified in the previous section, grayscaled and cropped. The experiment was performed by 93 observers, who carried it out on their personal computer and screen. Therefore, the viewing environment of each observer was uncontrolled, which could pose some challenges, mainly related to the resolution of the image, which will vary with the type of display and the distance of the observer from the screen. However, this effect is limited by the fact that Pavlovia automatically activates the full-screen view when the experiment starts. As the images were grayscale, a color calibration of the screen was not necessary. The display settings of each observer could have had an impact, although as discussed in Ref. 36, many studies have compared online behavioral experiments with lab-based ones, and they found that their data quality is usually equivalent.

The experiment was divided into 30 rounds. At each of them, an observer was presented a texture reference and 25 samples and was asked to select all the images that looked similar to the reference. No information other than this was provided to the observers and no definition of the words “similar” and “texture” was given before the experiment. The 26 textures, samples and reference, were selected randomly from the database, and therefore, rounds without instances of the reference image class in the 25 samples were possible. An example of an experiment round is shown in Fig. 8. Based on the results of this experiment, it is possible to evaluate how similar two texture images A and B are by defining a similarity coefficient $\mathrm{SIM}_{A,B}$ [Eq. (7)]. Here $n_{\mathrm{group}}(A,B)$ indicates the number of times that A and B have been grouped together, whereas $n_{\mathrm{appear}}(A,B)$ indicates the number of times they appeared together in the same screen, given that A is a reference image. Note that $\mathrm{SIM}_{A,B} \in [0,1]$:

$$\mathrm{SIM}_{A,B} = \frac{n_{\mathrm{group}}(A,B)}{n_{\mathrm{appear}}(A,B)}. \tag{7}$$

The results of this similarity evaluation process can be used to fill a matrix, as shown in Fig. 9. In this figure, we highlighted the boundaries between images belonging to different DTD classes. From this figure, it can be seen which classes are confused with each other, such as the “chequered” with the “grid,” the “flecked” with the “dotted” and the “knitted,” and the “scaly” with the “stratified”.



Trial n. 1/30

Click on all the images similar in appearance to the reference.

Press enter to continue.

Fig. 8 A screenshot of the psychophysical experiment performed.

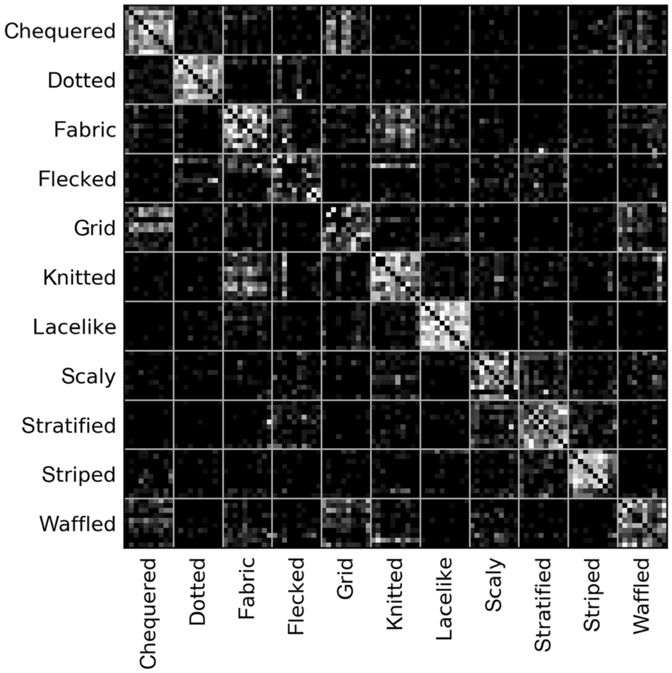


Fig. 9 The $\mathrm{SIM}_{A,B}$ matrix for the experiment performed.

If averaged over each class, this similarity coefficient matrix can be compared to the correlation matrix of the classification task 2, in Sec. 4.3. This is justified by the observation that, if the images belonging to a certain class are generally similar to those belonging to another according to the average human observer, it is more likely that the classification algorithm will confuse them. Therefore, we calculated the average confusion matrix of 1000 classification repetitions for the feature extraction techniques used in Sec. 4.3, and we then calculated the Spearman's rank correlation coefficient ρ between the two matrices after having collapsed them to one-dimensional vectors. We chose to use this measure because we want to evaluate the relationship

Table 6 Spearman's rank correlation coefficients between confusion matrix of the classification with different feature extraction methods and the $\text{SIM}_{A,B}$ matrix (Fig. 9).

Method	LBP	GLCM	HOG	Gabor	Wavelet
Spearman's ρ	0.51	0.49	0.47	0.48	0.38

between the two elements, without assumptions about its linearity or the type of distributions the data is obtained from. The results are shown in Table 6. The values of the ρ's show a moderate correlation and they seem to reflect the accuracies obtained in the second classification task. The only exception to this rule is the wavelet features, which show relatively low ρ but give a relatively high classification accuracy in the group. The same analysis can be performed with the mixed features. Without wavelet shrinkage, we obtain the results shown in Table 7. We have also conducted the experiment with wavelet shrinkage, resulting in ρ's on average 0.08 smaller than the ones without it. Interestingly, this shows how, while the best solution to classify images seems to be a mix of wavelet and LBP features vector, the results that better fit with the human observation are obtained by mixing wavelets with GLCM.

Another way to link this psychovisual similarity to the inspected features is by comparing the difference between the feature vectors of each pair of images with their similarity. This, too, can be gauged with the Spearman's ρ. First, we extracted the feature vectors as described in Sec. 4.1, calculated the distance between each of them, and compared the output with the similarity values. The results are provided in Table 8. Notice that they are all negative and quite small.

If we do the same for the mixed features, without wavelet shrinkage, we get results shown in Table 9. If we add wavelet shrinkage, the correlation coefficients are on average the same. The best possible choice of feature mixing is, in this case, the pure co-occurrence matrix features. On the other hand, wavelets appear to perform extremely poorly as stationary features. However, overall, the Spearman's ρ indicates that the correlation between the techniques used and the results of the psychovisual experiment performed is very weak. This is in line with the results of Sec. 4.3, which shows how demanding the DTD classification's task is.

Table 7 Spearman's rank correlation coefficients between confusion matrix of the classification with different stationarity-based mixed methods and the $\text{SIM}_{A,B}$ matrix (Fig. 9).

		Non-stationary				
		LBP	GLCM	HOG	Gabor	Wavelet
Stationary	LBP	0.51	0.47	0.45	0.46	0.49
	GLCM	0.53	0.49	0.52	0.51	0.53
	HOG	0.46	0.46	0.47	0.43	0.47
	Gabor	0.49	0.49	0.5	0.48	0.48
	Wavelet	0.38	0.34	0.34	0.35	0.38

Table 8 Spearman's rank correlation coefficients between feature vector distances and the $\text{SIM}_{A,B}$ matrix (Fig. 9).

Method	LBP	GLCM	HOG	Gabor	Wavelet
Spearman's ρ	−0.15	−0.18	−0.1	−0.15	−0.11

Table 9 Spearman's rank correlation coefficients between distances of feature vectors obtained with different stationarity-based mixed methods and the $\mathrm{SIM}_{A,B}$ matrix (Fig. 9).

		Non-stationary				
		LBP	GLCM	HOG	Gabor	Wavelet
Stationary	LBP	−0.15	−0.12	−0.15	−0.12	−0.13
	GLCM	−0.11	−0.18	−0.11	−0.15	−0.15
	HOG	−0.13	−0.11	−0.1	−0.11	−0.12
	Gabor	−0.16	−0.17	−0.16	−0.15	−0.13
	Wavelet	−0.11	−0.11	−0.11	−0.11	−0.11

6 Discussion and Conclusion

The results obtained in this work provide clues on how data stationarity is linked to human perception. First of all, the analysis of the chessboard images in Sec. 3 demonstrated its fundamental properties in relation to the simple case of a regular texture. According to the test used, the original chessboard image is data stationary, as are its stretched variations (Fig. 2). On the other hand, an increase of the chessboard patches dimension reduces the stationarity, particularly at lower scales and higher spatial frequencies (Fig. 5).

The classification experiments, described in Sec. 4, provide us with additional insights in relation to irregular textures. In the first task, discussed in Sec. 4.2, the p_j's, i.e., the p values resulting from testing a texture for stationarity at different scales, are used to mix the elements of GLCM and wavelet vectors. This, however, does not improve the classification accuracy. Additional analyses are reported in the Appendices. The results of this classification task suggest that Petrou and Sevilla's claim that while model-based texture features like GLCM are more suited for stationary images, frequency-based descriptors such as wavelet are preferred for non-stationary ones[3] has a limited validity in a classification framework. On the other hand, the second task, reported in Sec. 4.3, shows that using the stationarity information does improve the classification of the DTD images in their texture macrogroups. Compared to task 1, this assignment is more related to high-level texture perception.

Finally, the psychophysical experiment (Sec. 5) directly probes the link between perception and math. In its first part, it addresses the correlation between the confusion matrices of Sec. 4.3 and the similarity results, revealing how a mix of wavelet and LBP best replicates the average observer's response. It also demonstrates that the traditional texture features are very weakly correlated with the results of the visual experiment. Even in this context, despite the small size of the Spearman's ρ's, the $\mathbf{p}$-based mixing of features increases it.

In general, it is not clear if using wavelet shrinkage during the $\mathbf{p}$ calculation improves or reduces the relationship between visual and data stationarity. In some cases applying the wavelet shrinkage to $\mathbf{p}$ has a disruptive effect, whereas in some others, the effect is negligible. For example, in the chessboard experiment of Sec. 3, the shrinkage filters high-frequency artifacts, which degrade some $p_{\eta(j,l)}$'s of the $\mathbf{p}$, and it is, therefore, convenient. In the experiment of Appendix B, its application reduces the accuracy of the algorithm, whereas it increases it in Appendix E. Finally, it has a negative effect on the psychovisual analysis between mixed features and experimental similarity.

To conclude, this paper shows how stationarity information can be linked to the psychophysical attributes of a texture image and how evaluating it with LS2W processes can be used to improve a texture classification pipeline. There are various possible future steps that could better clarify the role of stationarity in texture. First, one can examine a wider variety of spatial stationarity metrics.[37–39] Even if the disadvantages of most of these have been highlighted in Ref. 8, their relationship with the perception of texture has yet to be assessed. The LS2W method itself can be improved. The scale analysis is currently performed at dyadic scales,[20] which allows fast extraction of the p of an image, but obtaining $\mathbf{p}$ on a continuous range of scales would provide more insight on its behavior. Another possible improvement to this approach is to expand it to color and spectral images and to find the best way to mix the various image channels. It has, in fact, been demonstrated that taking them into account increases the performance of texture analysis.[40] This has been already done for p in Ref. 21 but not for scale-dependent $\mathbf{p}$. Finally, p, $\mathbf{p}$, and $\mathbf{p}_j$ can be used to detect which image in a database has to go through a texture segmentation process.

7 Appendix A: Subimage Classification Using GLCM and Wavelet-Based Features at All Scales

In Sec. 4.2, we calculated the accuracy obtained by classifying images with f_{mix}, a feature vector obtained by mixing GLCM and wavelet elements. In this appendix, we compare the results of that experiment, shown in Table 2, to what can be achieved by simply combining the GLCM and wavelet vectors as in Eq. (8). Here c is an operation concatenating the vector $f_{ns,j}$ to $f_{s,j}$. With this, we achieve an accuracy of 72.3%, which is the best accuracy reached for this experiment (see Table 2):

$$f_{\text{comb},j}(f_s, f_{ns}) = c(f_{s,j}, f_{ns,j}). \tag{8}$$

8 Appendix B: Subimage Classification Using Variations of f_{mix}

For the classification based on f_{mix} (see Sec. 4.2), we also considered the cases in which $\mathbf{p}$ is calculated without wavelet shrinkage-based smoothing. Eckley et al.[17] demonstrated that the results of the stationarity test are more reproducible if smoothing is applied, but this could be counterproductive for the classification. Moreover, we assessed the case in which $f_s = \text{wavelet}$ and $f_{ns} = \text{GLCM}$ [see Eq. (6)] since this would provide us with experimental proof for the considerations proposed in Ref. 3, i.e., that the some techniques, like GLCM, are more suited to stationary images than others, such as wavelets.

The results obtained by calculating the p_j's of each subimage are shown in Table 10. We can see that the "rough" $\mathbf{p}_j$ that is obtained without applying wavelet shrinkage perform better. These outputs are partially in line with Petrou and Sevilla's[3] hypothesis discussed in Sec. 4.1, as using wavelet as stationary technique is worse than using GLCM. However, the accuracies reported in Table 10 are all smaller than those obtained with pure GLCM and wavelet features (see Table 2) and with a combination of the two (see Appendix A).

9 Appendix C: ALOT Analysis

As discussed in Sec. 1, we chose to use the DTD images in our analysis because of their vision-based arrangement. To provide an alternative, we also considered the Amsterdam Library of Textures (ALOT).[41] In particular, we studied the ALOT pictures mentioned in Ref. 6. In this paper, the authors classified a wide variety of existing texture databases according to certain

Table 10 Classification accuracy corresponding to the classification task 1 shown in Fig. 7, with various choices of stationary features and with wavelet shrinkage.

Shrinkage	Stationary	Non-stationary	Accuracy (%)
No	GLCM	Wavelet	64.62
No	Wavelet	GLCM	61.18
Yes	GLCM	Wavelet	61.81
Yes	Wavelet	GLCM	57.04

Table 11 Classification accuracy of task 1 performed on the ALOT images.

Method	GLCM	Wavelet	$f_{\text{mix,ws}}$	$f_{\text{mix},s}$
Average accuracy without shrinking (%)	72.18	65.99	70.46	71.92

characteristics, among which is texture stationarity, with explicit mention to the definition given in Ref. 3. Therefore, we applied the LS2W stationarity test (Sec. 2) to two datasets defined in this paper and extracted from the ALOT: one stationary ALOT-95-S-N and the non-stationary ALOT-40-NS-N. Setting the significance level α to 0.1 and the number of bootstrap iterations B to 10, only 40% of the ALOT-95-S-N images are classified as stationary by the test, whereas for the ALOT-40-NS-N group, this percentage is increased to 67.5%. This demonstrates the need for a common definition of data and visual stationarity.

Subsequently, we expanded the results reported in Sec. 4.2 by applying the first classification task to the ALOT. In particular, we merged the two classes ALOT-95-S-N and ALOT-40-NS-N. We adopted the same approach as Sec. 4.2, dividing each sample in 16 subimages of shape 256×256. The results of the classification are shown in Table 11, where $f_{\text{mix},ws}$ indicates the mixed features obtained without wavelet shrinkage and $f_{\text{mix},s}$ the ones with it. These results are similar to those attained in Sec. 4.2.

10 Appendix D: Dimension Dependence

In Sec. 4, we divided each image in subimages of dimension 32×32. This is in contrast with the results discussed in Ref. 8, whose experiment on the power assessment of the LS2W test shows that an image size of at least 128 pixels side is required to achieve good statistical power. However, this conclusion has been obtained based on artificial non-stationary models whose visual non-stationarity is extremely low [e.g., see Fig. 3d in Ref. 8]. Moreover, the chosen subimages' size is limited by the dimension of the images selected from the DTD.

In this appendix, we analyze how the size of the subimages can influence the results of Sec. 4.2. First, we divided the DTD images selected into bigger subimages. This has the drawback of reducing the total number of images available for the classification. We then repeated task 1 of the classification section. With subimages of size 64×64, which correspond to dividing the original picture into four squared sections, we obtain Table 12, and with subimages of size 42×42, we obtained Table 13.

As in Sec. 4.2, we reported the numbers obtained by mixing the features using the p_j's obtained without wavelet shrinkage, as applying it would slightly reduce the classification performance. One can see that the results obtained in Sec. 4.2 correspond, with minor variations, to the ones showed here.

Table 12 Classification accuracy of task 1 performed on images with size 64×64.

Method	GLCM	Wavelet	f_{mix}
Average accuracy (%)	63.55	58.84	57.62

Table 13 Classification accuracy of task 1 performed on images with size 42×42.

Method	GLCM	Wavelet	f_{mix}
Average accuracy (%)	68.77	67.15	62.32

Table 14 Classification accuracy of task 1 performed on images with size 128 × 128.

Method	GLCM	Wavelet	f_{mix}
Average accuracy (%)	95.26	95.01	93.19

As mentioned, Taylor et al.[8] suggested using square images with sides of at least 128 pixels. To satisfy this requirement without reducing the number of samples for the classification, we randomly extracted subpictures of size 128 × 128 from the selected DTD images. For each image, we derived 16 subpictures so that the number would correspond to the batch used in the calculation of Table 2, for a total of 1760 samples. The output of this experiment is shown in Table 14. In this case, the accuracy is strongly enhanced, probably due to the fact that it is likely that some of the classified pictures overlap. Nonetheless, the conclusions of Sec. 4.2 are still unaffected by the change of dimension of the images. In this case, mixing with wavelet shrinkage is the best performing method, and thus it is the number reported in Table 14.

11 Appendix E: Subimage Classification Using Image Source p

In Sec. 4.2, we extracted the mixed features vector f_{mix} using a different set of p_j's for each subimage. If we repeat the experiment with a common $\mathbf{p}_j$ for all subimages belonging to the same original image, we get the results shown in Table 15. Here we can see a clear improvement with respect to the case discussed in Appendix B, due to the fact that subimages with common origin have the same null terms. Nonetheless, the improvement is quite significant and it is interesting how the wavelet shrinkage further boosts it. In this case, the classification accuracy is actually improved in respect with the output obtained using pure features.

Table 15 Classification accuracy corresponding to the classification task 2 shown in Fig. 7, with common set of p_j's.

Shrinkage	Stationary	Non-stationary	Accuracy (%)
No	GLCM	Wavelet	77.79
No	Wavelet	GLCM	75.34
Yes	GLCM	Wavelet	79.96
Yes	Wavelet	GLCM	75.24

Acknowledgments

This work was partially supported by FRIPRO FRINATEK Metrological texture analysis for hyperspectral images (Project No. 274881) funded by the Research Council of Norway.

References

1. I. Florescu, *Probability and Stochastic Processes*, John Wiley & Sons, Inc., Hoboken, New Jersey (2014).
2. H. Tamura, S. Mori, and T. Yamawaki, "Textural features corresponding to visual perception," *IEEE Trans. Syst. Man Cybern.* **8**(6), 460–473 (1978).
3. M. Petrou and P. G. Sevilla, *Image Processing: Dealing with Texture*, John Wiley & Sons, Inc., Chichester, England (2006).
4. M. Petrou and S. I. Kamata, *Image Processing: Dealing with Texture*, John Wiley & Sons, Inc., Chichester, England (2021).

5. N. Bhushan, A. R. Rao, and G. L. Lohse, “The texture lexicon: Understanding the categorization of visual texture terms and their relationship to texture images,” *Cogn. Sci.* **21**, 219–246 (1997).
6. R. Bello-Cerezo et al., “Comparative evaluation of hand-crafted image descriptors vs. off-the-shelf CNN-based features for colour texture classification under ideal and realistic conditions,” *Appl. Sci.* **9**, 738 (2019).
7. M. Conni and H. Deborah, “Texture stationarity evaluation with local wavelet spectrum,” in *London Imaging Meeting*, pp. 24–27 (2020).
8. S. L. Taylor, I. A. Eckley, and M. A. Nunes, “A test of stationarity for textured images,” *Technometrics* **56**, 291–301 (2014).
9. E. V. Kurmyshev, M. Poterasu, and J. T. Guillen-Bonilla, “Image scale determination for optimal texture classification using coordinated clusters representation,” *Appl. Opt.* **46**, 1467–1476 (2007).
10. M. Cimpoi et al., “Describing textures in the wild,” in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3606–3613 (2014).
11. B. Julesz, “Experiments in the visual perception of texture,” *Sci. Am.* **232**, 34–43 (1975).
12. E. V. Kurmyshev and M. A. Cervantes, “A quasi-statistical approach to digital binary image representation,” *Rev. Mex. Fis.* **42**, 104–116 (1996).
13. B. Vidakovic, *Statistical Modeling by Wavelets*, John Wiley & Sons, New York (2009).
14. C. K. Chui, *An Introduction to Wavelets*, Academic Press, San Diego, CA (1992).
15. S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, CA (1999).
16. G. P. Nason, R. V. Sachs, and G. Kroisandt, “Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum,” *J. R. Stat. Soc. Ser. B* **62**, 271–292 (2000).
17. I. A. Eckley, G. P. Nason, and R. L. Treloar, “Locally stationary wavelet fields with application to the modelling and analysis of image texture,” *J. R. Stat. Soc. Ser. C* **59**, 595–616 (2010).
18. I. A. Eckley and G. P. Nason, “Efficient computation of the discrete autocorrelation wavelet inner product matrix,” *Stat. Comput.* **15**, 83–92 (2005).
19. A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge, UK (1997).
20. M. A. Nunes, S. L. Taylor, and I. A. Eckley, “A multiscale test of spatial stationarity for textured images in R,” *R J.* **6**, 20–30 (2014).
21. S. L. Taylor, I. A. Eckley, and M. A. Nunes, “Multivariate locally stationary 2D wavelet processes with application to colour texture analysis,” *Stat. Comput.* **27**, 1129–1143 (2017).
22. “Source of the imutils Python library,” https://github.com/jrosebr1/imutils (accessed November 2020).
23. “Describable textures dataset (DTD) website,” http://www.robots.ox.ac.uk/vgg/data/dtd/ (accessed September 2020).
24. “Documentation of the imread function in the OpenCV library,” https://docs.opencv.org/4.5.1/d4/da8/group__imgcodecs.html#imread (accessed March 2021).
25. F. Bianconi, R. Bello-Cerezo, and P. Napoletano, “Improved opponent color local binary patterns: an effective local image descriptor for color texture classification,” *J. Electron. Imaging* **27**, 011002 (2017).
26. R. M. Haralick, K. Shanmugam, and I. H. Dinstein, “Textural features for image classification,” *IEEE Trans. Syst. Man Cybern.* **SMC-3**, 610–621 (1973).
27. N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognit.*, Vol. 1, pp. 886–893 (2005).
28. T. Randen and J. H. Husøy, “Filtering for texture classification: a comparative study,” *IEEE Trans. Pattern Anal. Mach. Intell.* **21**, 291–310 (1999).
29. F. Bianconi and A. Fernández, “Evaluation of the effects of Gabor filter parameters on texture classification,” *Pattern Recognit.* **40**, 3325–3335 (2007).
30. K. He et al., “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
31. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv:1409.1556 (2014).

32. C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2818–2826 (2016).
33. G. Huang et al., "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 4700–4708 (2017).
34. J. Peirce et al., "Psychopy2: experiments in behavior made easy," *Behav. Res. Methods* **51**, 195–203 (2019).
35. "Pavlovia platform homepage," https://pavlovia.org/ (accessed June 2021).
36. M. Sauter, D. Draschkow, and W. Mack, "Building, hosting and recruiting: a brief introduction to running behavioral experiments online," *Brain Sci* **10**, 251 (2020).
37. A. Ephraty, J. Tabrikian, and H. Messer, "A test for spatial stationarity and applications," in *Proc. 8th Workshop Stat. Signal and Array Process.*, pp. 412–415 (1996).
38. S. Bose and A. Steinhardt, "Invariant tests for spatial stationarity using covariance structure," *IEEE Trans. Signal Process.* **44**, 1523–1533 (1996).
39. M. Fuentes, "A formal test for nonstationarity of spatial stochastic processes," *J. Multivariate Anal.* **96**, 30–54 (2005).
40. E. Cernadas et al., "Influence of normalization and color space to color texture classification," *Pattern Recognit.* **61**, 120–138 (2017).
41. G. J. Burghouts and J.-M. Geusebroek, "Material-specific adaptation of color invariant features," *Pattern Recognit. Lett.* **30**, 306–313 (2009).

Michele Conni received his BS and MS degrees in engineering physics from the Polytechnic University of Milan in 2015 with specialization in optics and photonics. He is currently studying for his PhD in computer science at Norwegian University of Science and Technology (NTNU), in collaboration with Barbieri Electronic, where he works in the research and development group.

Hilda Deborah received her BSc degree in computer science from the University of Indonesia in 2010, her MSc degree from Erasmus Mundus Color in Informatics and Media Technology in 2013, and her PhD in computer science from NTNU and the University of Poitiers in 2016. She is a researcher at NTNU. Her current research interests are hyperspectral imaging and texture analysis.

Peter Nussbaum received his MSc degree from the Colour and Imaging Institute, University of Derby, Derby, UK, in 2002 and his PhD in imaging science from the University of Oslo, Oslo, Norway, in 2011. He is an associate professor of color imaging at the Colour and Visual Computing Laboratory, NTNU, Gjøvik, Norway.

Phil Green received his MSc degree from the University of Surrey in 1995 and his PhD from the former Colour and Imaging Institute, University of Derby, Derby, UK, in 2003. He is a professor of color imaging at the Colour and Visual Computing Laboratory of NTNU. He is also a technical secretary of the International Color Consortium, the body that standardizes the ICC profile format and promotes color management internationally.