Perceptual Texture Similarity Estimation: An Evaluation of Computational Features

Xinghui Dong, Junyu Dong, and Mike J. Chantler

Abstract—Estimation of texture similarity is fundamental to many material recognition tasks. This study uses fine-grained human perceptual similarity ground-truth to provide a comprehensive evaluation of 51 texture feature sets. We conduct two types of evaluation and both show that these features do not estimate similarity well when compared against human agreement rates, but that performances are improved when the features are combined using a Random Forest. Using a simple two-stage statistical model we show that few of the features capture long-range aperiodic relationships. We perform two psychophysical experiments which indicate that long-range interactions do provide humans with important cues for estimating texture similarity. This motivates an extension of the study to include Convolutional Neural Networks (CNNs) as they enable arbitrary features of large spatial extent to be learnt. Our conclusions derived from the use of two *pre-trained* CNNs are: that the large spatial extent exploited by the networks' top convolutional and first fully-connected layers, together with the use of large numbers of filters, confers significant advantage for estimation of perceptual texture similarity.

Index Terms—Evaluation, features, perceptual similarity, similarity measures, texture similarity

1 Introduction

Texture has been an extremely popular subject in computer vision over the last fifty years [56], [57], [75], [82] but surprisingly, the task of texture similarity estimation, as judged by human observers, has not received as much attention as classification [91] or segmentation [61].

It is often characterised as "a spatial organisation of a set of basic elements or primitives" [63]. Recently, "orderless" techniques [118], [134] have been successfully exploited in Convolutional Neural Networks (CNNs) in order to remove higher-level positional data, however, their *convolutional* layers still compute local spatial statistics. A wide variety of texture features have thus been designed, or learnt, that encode different types of spatial statistics. These tend to be either of 2nd- or higher order, as 1st-order statistics do not encode pixel location information.

Examples of popular features that capture 2nd-order statistics include those based on the biologically motivated Linear-Nonlinear-Linear (LNL) and similar models [48], [70], [81]. These have evolved from small sets of hand-crafted filters [71] to being the basis for the CNNs that are so successful today [24], [108]. Naturally if the positional data of the local filters is retained (as performed implicitly in CNNs), or they are locally combined in some way (e.g., quadrature filters [99]) then phase information can be exploited.

In addition, there are many approaches that are designed to capture complex local patterns in which their local higher-order statistics (HOS) and phase data are critical, e.g., [92]. These include features explicitly designed to analyse regular and near-regular textures [73], [74], [76], [78] in which wallpaper symmetries have shown to be an important part of human perception [68]. Others include the trace transform [64], co-occurrence matrices [57] and texton-based approaches [72], [122].

In general, therefore, existing texture features estimate local and global 2nd-order statistics, or *local* HOS. However, we have come across few that utilise longer-range HOS, which are known to be used by the human visual system (HVS) [46], [98], [113]. An exception is CNNs as their fully-connected layers have access to positional data.

Hence the goals of this paper are to analyse the spatial extent and statistical nature of existing feature sets, to investigate their performance using human-derived similarity data, and compare their performance with CNNs.

1.1 Related Work

As stated above there is no agreement as to a formal definition of texture [56], [60] however, the term has been extensively used in a number of research disciplines.

Human vision researchers have used it to describe a wide variety of stimuli used in both psychophysical and neuroscience studies. They vary from the seemly simplistic binary textures [46], [63] through to more complex images that explore characteristics such as symmetry [30], [68] and HOS [42]. The former for instance have been used to investigate dipole statistics and Julesz's famous conjecture hypothesised that humans cannot distinguish between textures with identical 2nd-order statistics. He later proved this false and

- *X. Dong is with the Centre for Imaging Sciences, The University of Manchester, M13 9PT United Kingdom. E-mail: dongxinghui@gmail.com.*
- *J. Dong is with the Department of Computer Science, Ocean University of China, Qingdao, Shandong 266071, China. E-mail: dongjunyu@ouc.edu.cn.*
- *M. J. Chantler is with the Texture Lab, School of Mathematical and Computer Sciences, Heriot-Watt University, EH14 4AS Edinburgh, United Kingdom. E-mail: m.j.chantler@hw.ac.uk.*

Manuscript received 7 Oct. 2018; revised 13 Nov. 2019; accepted 9 Dec. 2019. Date of publication 7 Jan. 2020; date of current version 3 June 2021. (Corresponding author: Xinghui Dong). Recommended for acceptance by Y. Liu. Digital Object Identifier no. 10.1109/TPAMI.2020.2964533

0162-8828 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.

went on to develop a "texton" based approach in order to identify discriminatory features [63]. In addition, natural texture images, e.g., Brodatz's [19], have been used extensively for studies of human perception of texture properties [27], [60], [103].

In contrast computer vision researchers have developed many feature sets and systems designed to perform such tasks as segmentation, classification and retrieval. They often characterise texture as referring to the spatial organisation of a set of primitives [33], [75], [117], [121] similar to Juesz [63] in many ways. This idea of a set of texture primitives coupled with placement rules has been extended by Liu and colleagues to cover *near-regular* textures in which lattice-based geometric, colour, and lighting distortions can be captured (and synthesised) in order to characterise texture [74], [76], [78].

1.1.1 Previous Reviews and Surveys

While there have been numerous papers published proposing new texture features, surveys that have provided wide reviews of such features are scarcer.

Haralick [56] provided one of the earliest surveys of models of feature extraction, dividing them into structural and statistical approaches. Van Gool [121] used the same classification to survey feature measures for texture segmentation, while Reed and du Buf [104] reviewed a large number of approaches specifically for unsupervised segmentation. Tuceryan and Jain [117] examined geometric, random field, fractal, and signal processing models for the tasks of segmentation and classification. Ojala *et al.* [90] compared features for classification using their distributions, while Randen and Husøy [102] evaluated many filter-based features using a pixel-wise texture classification task. Xie and Mirmehdi [130] proposed a taxonomy of texture measures and used this to review a "galaxy" of features, and Dana [33] summarised traditional and deep learning based computational models used for four texture tasks. Most recently Liu *et al.* [75] surveyed a large number of feature sets focusing on those presented over the last 20 years and suited to the purposes of classification, segmentation, synthesis, and shape from texture.

Of these studies, none evaluate features explicitly for the purposes of texture similarity estimation. This may well be due to the lack of texture databases with associated sets of perceptual texture similarity data.

1.1.2 Evaluation against Perceptual Judgements

Human perceptual judgements have been used as ground-truth to evaluate a wide variety of computer algorithms. For instance, observers' rankings have been used to evaluate search engine performance [13], [58], [87]; while Eitz *et al.* [41] used a 7-point Likert scale to gather perceived similarity between sketches and images, and Zhang and colleagues employed 2AFC and JND assessments of *distortions* of images caused by six types of artefact [133]. In addition, large numbers of human judgements have also been used for assessing bidirectional texture functions [47], computational image saliency [16], [116] and visual realism [44].

However, the direct use of human studies for evaluation of texture features is less common. This is particularly the case when we consider research that employs *fine-grained judgements*[1] using a larger number of response alternatives, e.g., a 7-point Likert scale rather than the more common *binary* sets of categories using, for instance, "similar" and "dissimilar". Lin *et al.* [73] examined the correlation between observers' ratings judged on a 4-point scale of the quality of near-regular texture synthesis [78]. Cimpoi *et al.* [28] used Amazon Mechanical Turk to label 5,640 texture images with multiple perceptual properties, while Sharan *et al.* [107] asked mTurkers to label distorted images with material categories. Matthews *et al.* [84] asked observers to perform two-alternative pair-wise comparisons of *specific* texture qualities. Zujovic [136] developed a "Visual Similarity by Progressive Grouping" (ViSiProG) procedure to generate similarity matrices from which they extracted disjoint clusters, while Pappas *et al.* [95] used the same clusters to analyse the performance of nine similarity metrics (focusing on structural types). However, they only used the (binary) intra-cluster similarity. Payne *et al.* [97] and Santini and Jain [105] compared human and computational rankings of 100 textures.

Other researchers have sought to develop features that estimate particular perceptual texture properties. For instance Tamura *et al.* [114] used pair-wise comparison of six perceptual properties to estimate interval scales which they used to assess computational measures. Other authors have also used these data to perform similar assessments [50] or have used ranking methods to assess the effectiveness of texture features designed to measure perceptual properties [8], [12].

Perceptual texture dimensions have been researched by Cho *et al.* [27], Rao and Lohse [103], and Heaps and Handel [60] with the latter concluding that texture similarity is context dependent, and that a dimensional model is not appropriate.

Finally, while Clarke *et al.* [29] did not evaluate computational features against perceptual judgements *per se*, they did compare two sets of fine-grained perceptual similarity data with each other. They collected 1000 pairs-of-pairs judgements and compared these against pair-wise similarities for 334 *Pertex* [3] textures estimated using Isomap dimensionality reduction [115] applied to the results of free-grouping by 30 human observers [54] (see Section 3.3.2).

Thus, while a number of studies have assessed the ability of features and measures to estimate human-perceived texture similarity, none of the papers described above use such data to compare more than a handful of feature sets at a time. An exception to this is provided by two short papers by Dong *et al.* [37], [38]. The contribution here, beyond [37] and [38], is described below.

1.2 Contribution of This Survey

In comparison to the papers reviewed above, and compared to [37] and [38], this paper adds an in-depth survey of the 51 texture feature sets and a detailed analysis of their spatial extent and order of statistics used. It evaluates the results more extensively and provides additional significance testing. It describes the ground-truth data and evaluation methods more thoroughly, and it investigates the importance of long-range interactions in more detail. In addition, it examines the

[1] By "fine-grained" we simply mean the case where three or more categories (or ordinal or ratio data) are used to represent pair-wise similarity.

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.

performance of Random Forest regressors [18] and adds an evaluation of CNNs using two types of neural layers, convolutional and fully-connected, obtained from two pre-trained networks.

The contributions are therefore as follows. (1) We analyse 51 traditional feature sets in detail in terms of their statistical nature and spatial extent. (2) We show that a multi-resolution approach is significantly better than using single resolutions for these features. (3) We use two types of "fine-grained", ordinal similarity human ground-truth in two separate evaluation protocols. (4) We provide a detailed description of the use of the Block Randomised experiment to investigate the effect of long-range interactions on human perception and show that such interactions are important in human similarity judgements. (5) We show that the combination of all 51 feature sets in a Random Forest regressor [18] produces significantly better results than any individual feature set. (6) Finally, we show that features derived from two different pre-trained CNNs outperform the traditional features, including the case when they are combined using a Random Forest.

1.3 Organisation of This Survey

The computational texture features are reviewed in Section 2 while Section 3 describes the selection of the texture database and its ground-truth. Section 4 introduces two different evaluation protocols corresponding to the two experiments reported in Sections 5 and 6. Section 7 investigates the importance of long-range interactions to human texture perception, CNN features are assessed in Section 8, and conclusions are presented in Section 9.

2 Computational Texture Features

Many computational texture features have been proposed over the last fifty years [57], [75], [104], [117], [121], [130]. However, none have provided a uniform treatment of features based on the order and spatial extent of their statistical properties. Our motivation in providing such a survey here is: (1) 1st-order statistics do not encode spatial relationships and can therefore be conveniently ignored; (2) 2nd-order statistics, particularly in the form of the autocorrelation function or power spectrum, are well understood and often used to represent periodic and regular textures; and (3) complex aperiodic structures can be encoded in higher order statistics (HOS) [93] (including phase spectra [94]) and in our experience provide significant challenges to computational similarity estimation. We therefore use a simple two-stage statistical model to provide a uniform means of comparison of the feature sets examined here.

2.1 A Two-Stage Statistical Model

For texture classification [91] and retrieval [82], feature extraction is often implemented in two stages: the first (Stage I) is used to compute local statistics, while the second (Stage II) often aggregates these to provide global, image-wide features. We use this simple two-stage model to simplify analysis of features' statistical properties and spatial supports. This viewpoint does not limit our choice to feature sets designed purely from a statistical viewpoint, nor prevent us from employing other categorisation methods.

Note that where intermediate values are calculated we merge such processing into Stage II. In the case of segmentation algorithms [61] we compute simple global, Stage II, features from their feature maps.

2.2 Long-Range Interactions

Throughout this paper we refer "long-range" interactions. By this we mean 2nd- and higher-order pixel dependencies that occur over image distances of 20 pixels or more. We have chosen this value as many texture features do not compute HOS at such ranges. In contrast, 2nd-order statistics such as those encoded in the autocorrelation function can be used to capture image-wide, periodic, regular patterns at relatively low cost. Periodicities can be further classified using 17 different "Wallpaper Groups" that can be thought of as using placement rules to position "tiles", where the rules are generated from combinations of four basic operations: translation, rotation, reflection and glide-reflection [30], [68], [78]. "Semi-regular" versions of these patterns can be generated (or analysed) by either deforming the underlying lattice or introducing intensity (colour) variations to corresponding pixels across tiles [78]. Liu and colleagues developed the G-A Score, derived with human assistance for lattice generation [76], which neatly quantifies such deviations from regular periodic patterns and therefore provides two measures of long-range HOS.

2.3 The Feature Sets

A feature set was selected for investigation if (1) it is popular in the literature; (2) its source code is published or it can be straightforwardly implemented according to the original publication; (3) the features can be automatically extracted without manual assistance; and (4) it can be applied to any texture. We ignored variants that were reported to produce similar results. We did not include any *global* phase features as we believe that phase unwrapping is still an open problem [131]. However, we did include *local* phase based methods, including Joint Statistics of Complex Wavelet (JSCW) [99] and Local Phase Quantisation (LPQ) [92].

The above provided 46 feature sets. To these we added three Canny [22] and two further Sobel [112] feature sets to complement GMAGGDIRSOBEL [90]. In total 51 feature sets were investigated. We refer to these as "conventional" feature sets. Random Forest regressors and CNN implementations are discussed in Sections 6.7 and 8.

2.3.1 Feature Set Coverage

In order to provide insight as to the coverage and diversity of the feature sets selected, we have categorised these according to an existing taxonomy. Tuceryan and Jain [117] and Xie and Mirmehdi [130] both provided eloquent categorisations and here we use the four categories proposed in [130]. These comprise *signal processing based*, *statistical*, *structural* and *model-based* approaches. Furthermore, we have additionally divided the features into histogram and non-histogram based approaches (Table 1 parts (a) and (b)).

2.3.2 Signal Processing Based Features (◆)

These features are often obtained using the local "energy" (e.g., variance estimation) of linear filter responses. They

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.

TABLE 1
Summary of 51 Computational Texture Feature Sets: (a) Histogram-Based and (b) Non-Histogram-Based

Identifier	Full Name	Year	Ref.	Tasks	Categ.	Statistical Property	Effective Spatial Extent
SAC	Centre-Symmetric Auto-correlation	1995	[59]	C	♠	2nd	3×3
SRAC	Centre-Symmetric Rank-Order Auto-correlation	1995	[59]	C	♠	2nd	3×3
SVR	Centre-Symmetric Variance Ratio	1995	[59]	C	♠	2nd	3×3
VAR	Rotation Invariant Local Variances	2002	[91]	C	♠	Higher	5×5 (Radius = 2)
GDIRCANNY	Canny Gradient Direction Distributions	N/A	N/A	PC	♥	Higher	9×9
GDIRSOBEL	Sobel Gradient Direction Distributions	N/A	N/A	PC	♥	Higher	3×3
GMAGCANNY	Canny Gradient Magnitude Distributions	N/A	N/A	PC	♥	Higher	9×9
GMAGGDIRCANNY	Joint Distributions of Canny GMAG and GDIR	N/A	N/A	PC	♥	Higher	9×9
GMAGGDIRSOBEL	Joint Distributions of Sobel GMAG and GDIR	1996	[90]	C	♥	Higher	3×3
GMAGSOBEL	Sobel Gradient Magnitude Distributions	N/A	N/A	PC	♥	Higher	3×3
LBPBASIC	Basic Local Binary Patterns (LBP)	2009	[11]	C	♥	Higher	3×3
LBPDF	Local Derivative Filters Based LBP	2009	[11]	C	♥	Higher	3×3
LBPHF	Local Binary Pattern Histogram Fourier	2009	[10]	C	♥	Higher	5×5 (Radius = 2)
LBPRIU2	Rotation-Invariant Uniform LBP	2002	[91]	C	♥	Higher	5×5 (Radius = 2)
LBPRIU2&VAR	Joint Distributions of LBPRIU2 and VAR	2002	[91]	C	♥	Higher	5×5 (Radius = 2)
LDP	Local Derivative Patterns	2010	[132]	C	♥	Higher	3×3
LDPSE	Spatially Enhanced LDP	2010	[132]	C	♥	Higher	3×3
RI-LPQ	Rotation-Invariant Local Phase Quantisation	2008	[92]	C	♥	Higher	9×9
VZ-MR8	Varma & Zisserman's MR8 Textons	2005	[122]	C	♥	Higher	5×5 (Effective Filter Size)
VZ-MRF	Varma & Zisserman's Markov Random Field Textons	2009	[123]	C	♥	Higher	19×19
VZ-NBRHD	Varma & Zisserman's Neighbourhood Textons	2009	[123]	C	♥	Higher	19×19

(a) Histogram-Based

Identifier	Full Name	Year	Ref.	Tasks	Categ.	Statistical Property	Effective Spatial Extent
DCT	Discrete Cosine Transform Based Channel Filters	1992	[89]	S	◆	2nd	*
EIGENFILTER	Eigen Filters	1983	[9]	S	◆	2nd	*
GABORBOVIK	Bovik's Localised Gabor Filters	1990	[17]	S	◆	2nd	*
GABORENERGY	Gabor Energy Filters	1989	[48]	S	◆	2nd	*
GABORJFFD	Dyadic Gabor Filter Bank (Frequency Domain)	1991	[61]	S	◆	2nd	*
GABORJFSD	Dyadic Gabor Filter Bank (Spatial Domain)	1991	[61]	S	◆	2nd	*
GABORMM	Manjunath & Ma's Gabor Wavelet Filter Bank	1996	[82]	R	◆	2nd	*
JSCW	Joint Statistics of Complex Wavelet	2000	[99]	PC	◆	Higher	17×17
LAWS	Laws' Masks	1980	[71]	S	◆	2nd	*
LM	Leung & Malik's Filter Set (Bank)	2001	[72]	PS	◆	2nd	*
MR8	Maximum Response Filter Set	2005	[122]	PS	◆	2nd	*
RFS	Root Filter Set	2005	[122]	PS	◆	2nd	*
RING & WEDGE	Ring and Wedge Filters	1985	[31]	S&C	◆	2nd	*
S	Schmid's Filter Set	2001	[106]	PS	◆	2nd	*
ACF	Autocorrelation Functions	2003	[50]	R	♠	2nd	*
CVM	Covariance Matrices	1996	[77]	S	♠	2nd	*
GLADH	Absolute Grey Level Difference Histograms	1976	[129]	C	♠	2nd	9×1
GLCM	Grey Level Co-occurrence Matrices	1973	[57]	C	♠	2nd	9×1
GLSDH	Signed Grey Level Difference Histograms	1986	[119]	C	♠	2nd	9×1
GLSDSH	Signed Grey Level Difference and Sum Histograms	1986	[119]	C	♠	2nd	9×1
GLSH	Grey Level Sum Histograms	1986	[119]	C	♠	2nd	9×1
GLGLM	Grey Level Gap Length Matrices	1994	[125]	PC	♠	Higher	Length of Longest Gap
GLH	Grey Level Histogram	2009	[130]	PC	♠	1st	*
GLRLM	Grey Level Run Length Matrices	1975	[51]	C	♠	Higher	Length of Longest Run
MSA	Multi-scale Autoconvolution	2005	[101]	C	♠	3rd	*
SRDM	Surrounding Region Dependence Method	1999	[66]	C	♠	2nd	7×7
TT	The Trace Transform	2001	[64]	C	♠	2nd	Length of Longest Trace Line
FRACTALDIMENSION	Fractal Dimension	1993	[25]	S	♣	2nd	17×17
GMRF	Gaussian Markov Random Field	1985	[26]	C	♣	Higher	5×3
MRSAR	Multi-resolution Simultaneous Autoregressive	1992	[83]	S&C	♣	Higher	19×19

(b) Non-Histogram-Based

(1) "S", "C", and "R" denote segmentation, classification and retrieval tasks; (2) "PS" and "PC": these feature sets can potentially be used for segmentation and classification tasks respectively; (3) ◆, ♠, ♥ and ♣: signal processing based, statistical, structural and model-based features; (4) **: these feature sets work over the whole image; (5)* M *and* N*: the height and width of an image; and (6) bold fonts (in Column "Identifier") indicate feature sets which have been revised compared with the original algorithm.*

have many names, e.g., LNL, FRF (Filter-Rectify-Filter), etc., and have been used by both computer vision and vision science communities [70]. They map onto our two-stage model: the linear filtering process is regarded as the first stage while global feature extraction, typically variance estimation is treated as the second stage.

The majority of these features are designed and implemented in either the frequency or spatial domains. Such filters include: eigenfilters [9]; discrete cosine transform [89]; Laws' masks [71]; and the Gabor (wavelet) filters [17], [48], [61], [82]. Filters can also be used together in order to encode various textures [122], for instance, the Ring and Wedge filters [31], the LM filter bank [72], the S filter bank [106], and the RFS or MR8 filter banks [122]. In addition, filters used in a quadrature configuration, e.g., Joint Statistics of Complex Wavelet (JSCW) [99], are designed to extract local phase information.

Providing that (1) the filters are linear, and (2) the "energy" estimator computes variance, then the features can be transformed from the frequency domain to the spatial domain (and back) using Parseval's theorem [96]. This allows us to (1) compare feature sets in a single domain in order to analyse spatial extent, and (2) it enables us to

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.

conclude that many of the signal processing based features that we have examined (excepting JSCW) only utilise 2nd-order, power spectrum information and do not exploit HOS (see Table 1). Note that we only used the power spectra of the response matrices obtained using Localised Gabor Filters (GABORBOVIK) [17]. For all signal processing based features, excepting JSCW [99], the original post-processing on response matrices (normally comprising *local* variance estimation) was discarded and replaced by a *global* variance estimator (the square operation was applied to each response matrix and the mean was computed from each squared response matrix across the whole image).

Thus, due to the fact that the power spectrum cannot retain aperiodic image structure [94], the majority of the features discussed here are only able to capture *periodic* patterns. An exception is JSCW [99] which uses quadrature filters to estimate local phase information. However, the spatial extent of JSCW are 17×17 pixels, and thus they do not encode *long-range* aperiodic image structure either.

2.3.3 Statistical Features (♠)

1st-, 2nd- and higher order statistical features have been used to describe the distribution of grey levels. Popular 1st-order statistics include the mean of grey levels and the grey level histogram but it should be noted that such statistics do not capture spatial relationships.

In contrast 2nd-order features characterise the relationship between pairs of pixels defined by their relative position. Grey level co-occurrence matrices (GLCM) [57] provide one of the most classical feature sets of this type. Another similar approach is the use of absolute grey level difference histograms (GLADH) [129]. The histograms of the signed grey level differences (GLSDH), the grey level sum (GLSH) and the combination of these (GLSDSH) were further investigated by Unser [119]. Kim *et al.* [66] also designed a surrounding region dependence method (SRDM). Perceptual texture properties were explicitly modelled using the autocorrelation function (ACF) in [50] which, by the Wiener-Khinchin theorem, is known to be directly related to the power spectrum. Trace transform (TT) features [64] generally only compute 1st- and 2nd-order statistics on trace lines. Harwooda *et al.* [59] introduced local centre-symmetric covariance feature sets, including two local centre-symmetric auto-correlations with linear and rank-order versions (SAC and SRAC), a related covariance measure (SCOV) and a variance ratio (SVR). Ojala *et al.* [91] augmented Local Binary Patterns (LBP) with local variance estimators (VAR). A local covariance matrix based feature set (CVM) was also proposed by Liu and Madiraju [77]. Differing the original covariance matrix features, the mean and standard deviation were computed from each regional descriptor matrix and were combined into a feature vector.

Higher order statistics characterise pixel relationships between *three* or more pixels. For instance grey level run length matrix (GLRLM) [51] and grey level gap length matrix (GLGLM) [125], encode more complex spatial patterns. Rahtu *et al.* [101] proposed an affine invariant image transform, i.e., multi-scale autoconvolution (MSA), which calculates statistics based on point triplets.

Generally, computing 2nd or higher order statistics that capture information *additional* to the ACF is expensive. Hence these types of features are not often calculated over larger spatial support (see Table 1). Thus, it is rare to have processing that encodes long-range aperiodic image structure incorporated into these features.

2.3.4 Structural Features (♥)

Structural features generally assume that textures comprise *spatial primitives* (e.g., textons [63]) that are placed according to *spatial placement rules* [56], [124].

Julesz introduced the concept of textons [63] which has been incorporated into the design of many features [72], [106], [122], [123] and is used in "Bag-of-Words" (BoW) techniques, e.g., [109]. The occurrence frequency of textons is often exploited by such approaches and thus histogram comparison is often used with local structural features. Thus similar use of gradient magnitudes and directions [90]; local binary patterns (LBP) [10], [11], [91]; local derivatives [132]; and local phase information [92]; can all be considered as using the same basic approach as texton-based methods. In addition for gradient-based features, joint distributions (edge direction *and* magnitude) are often used. For example for the Sobel [112] operator (GMAGGDIRSOBEL) [90] we extracted histograms of gradient magnitudes (GMAG), and gradient directions (GDIR) for both Canny [22] and Sobel [112] edge detectors. In addition, we also used the joint distributions for Canny (GMAGGDIRCANNY). See Table 1a. (Note that these gradient-based feature sets can also be classified as signal processing as they use filter responses).

As regards the two-stage model, we consider Stage I to comprise texton generation and labelling, and Stage II to cover histogram generation. Thus, 2nd- and higher order statistics are only captured in Stage I which typically only operates over small spatial extents. Thus, higher order statistics are only calculated on relatively small local regions and hence, these structural features cannot encode long-range aperiodic image information.

2.3.5 Model-Based Features (♣)

Spatial process models' parameters are often used to describe textures. We have selected fractals [25], multi-resolution simultaneous autoregressive (MRSAR) models [83] and Markov Random Fields (MRF) [26] here, due to their popularity and availability.

For fractals we use the implementation provided by Smith and Burns [111] and computed variances of four fractal dimension estimates as the features. MRF were implemented as in [26], while for MRSAR we used its non-rotation-invariant version [83]. Due to the computational complexity of the latter the largest spatial extent that we employed was 19 × 19 pixels. The mean and standard deviation were computed from the model coefficients and error matrices estimated at each neighbourhood level and combined into a single feature vector.

For the three model-based feature sets used in this research, 2nd- and higher order statistics are only computed on relatively small local neighbourhoods and thus, once again, they cannot be used to capture long-range aperiodic image structure.

2.3.6 Feature Implementation Notes

We used the original source code for the features listed in Table 1 as far as practicable. If the source code was not

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.

available, we used the implementations by other authors or implemented it according to the publication. Parameters were set to those reported for optimal performance. Further tuning the parameters of a feature set or feature selection was avoided as much as possible.

2.4 Texture Measures Based on CNNs

In contrast to the conventional features described above, the majority of which have been explicitly designed by their proposers, CNNs learn their filters directly from data. Several of these *pre-trained* CNNs have been used to estimate similarity for texture synthesis and image generation. Gatys *et al.* [52] used Gram matrices computed from the convolutional layers of a CNN [108] as texture descriptors. Texture generation was performed by maximising the similarity between the Gram matrices of the original image and those of the generated image. In addition, Ustyuzhaninov *et al.* [120] used this type of similarity data to assess the quality of synthesised textures. Dosovitskiy and Brox [39] further introduced a set of loss functions based on the similarity calculated between CNN features which were extracted from the original and generated images. Bergmann *et al.* [15] built on Gatys' work to tackle periodic textures using generative adversarial networks (GANs) that provide manifolds on which texture distances can be measured, however, they did note that the approach can drop modes. Cimpoi *et al.* [28] used VGG-M [24] and VGG-VD-19 [108] models for recognition of textures in cluttered environments; found that this approach produced state-of-the-art results; and that CNN features transfer across domains without the need for adaptation.

We have therefore used these pre-trained CNNs (VGG-M [24] and VGG-VD-19 [108]) and compared their performances against the 51 "conventional" feature sets described above. The results for these networks, in a variety of configurations, are separately described in Section 8.

2.5 Summary of Selected Texture Feature Sets

Table 1 summarises feature sets in terms of tasks, category, the highest order statistical property, and the maximal effective spatial extent (used for computing HOS). With respect to these feature sets we found that: (1) the signal processing features examined do not exploit the phase data but only use power spectra, with the exception of JSCW [99] which calculates higher order statistics on *local* neighbourhoods; and (2) the statistical, structural and model-based features, excepting three: GLGLM [125], GLRLM [51], and MSA [101], do not compute higher order statistics at long ranges, i.e., $> 19 \times 19$ pixels.

In general, long-range periodic image structure, e.g., regular or near-regular texture [76], can be modelled using image-wide 2nd-order statistics (e.g., power spectra) or long-range HOS while long-range aperiodic image structure can only be encoded using long-range HOS, and hence only a few of the 51 feature sets considered encode these data.

We note that the G-A Score [76] which has been used extensively for near-regular texture synthesis and recurring pattern discoveries, captures longer-range HOS. However, we have not selected it for inclusion here due to the challenges of developing an auto-lattice detection algorithm for textures that do not have obvious lattice, or near-regular structures. See Fig. 10, rightmost images of rows 1, 3 and 4 for examples (rows numbered from top).

TABLE 2
Summary of 12 Published Texture Databases

Texture Database	Number of Textures	Constant Illumination	Constant Viewpoint	Perceptual Similarity Available
Brodatz [19]	112	×	×	√
CUReT [34]	61	√	√	×
KTH-TIPS [4]	10	√	√	×
KTH-TIPS2 [4]	11	√	√	×
Meastex [111]	69	×	×	×
Outex [91]	320	√	√	×
Pertex [3]	334	√	√	√
PhoTex [2]	64	√	√	×
RegTex [73]	43	×	×	×
STex [5]	476	×	×	×
UIUCTex [6]	25	×	√	×
VisTex [1]	167	×	×	×

3 Selecting Database and Ground-Truth

In this section we present the motivations for selecting the database used here, and for completeness we describe how the associated sets of ground-truth were derived.

3.1 Texture Database Selection

We identified three criteria for selection of the database. The first two concern illumination and viewpoint conditions as variation in either can affect human perception and the outputs of texture features. For instance directional illumination of Lambertian surface textures is known to act as a directional filter of the ACF [23]. We therefore wanted to remove these possible confounding factors and required texture images used to be acquired under constant viewpoint and constant illumination. Naturally, for real-world applications both of these factors should be controlled or accounted for [23], [34], [88].

Our third criterion concerned ground-truth, i.e., the availability of "human-judged texture similarity". That is the judgement of a typical observer, $s_{a,b}$, as to the similarity of textures a, and b. We would like such data to be fine-grained with ($s_{a,b} \in \mathbb{Q}$ *or* $\mathbb{R}$), as finer-grained similarity data provides a *more discriminating assessment*.

Table 2 summarises 12 published texture databases according to the number of textures and our three criteria (for more details, please refer to the supplementary material). It can be seen that only the *Pertex* database [3] satisfies all three criteria and thus it was selected for use.

The surprising result exposed in Table 2 is that only the *Brodatz* [19] and *Pertex* [3] databases are available with fine-grained ground-truth. Hence before describing the *Pertex* database we first explore below why only binary similarity data are commonly available with texture databases.

3.1.1 *Characteristics of Common Ground-Truth*

The majority of computer vision texture databases have been collected for testing classification [91], segmentation [17], [61] and retrieval [65], [82] algorithms.

For texture classification assessment, ground-truth is often provided as class labels, and so only binary similarity, $s_{a,b} \in \{same\ label, different\ label\}$ can be obtained. Similarly, for

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.

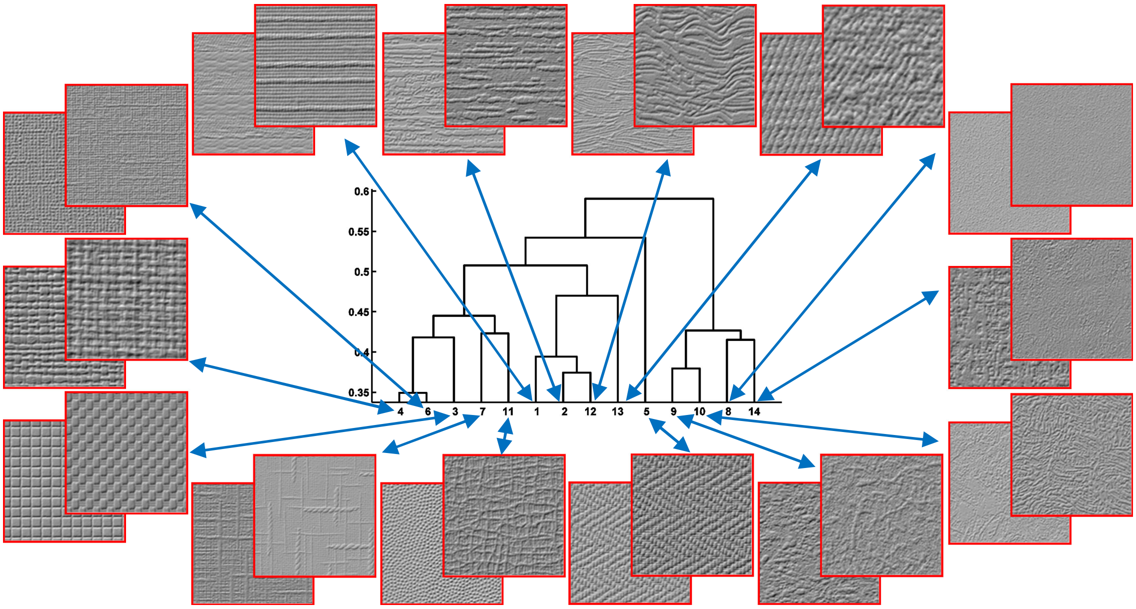


Fig. 1. Dendrogram (cut at 0.337) obtained from the 8D-ISO similarity matrix [29], along with two representative textures [3] of each cluster.

segmentation, the data that can be derived is also binary: $s_{a,b} \in \{same\ region, \ different\ region\}$. Texture retrieval experiments on the other hand often use non-overlapping sub-images derived from perceptually homogeneous parent images. Retrieval performance is based upon how many of a query's siblings are retrieved. Thus again, the similarity data that can be derived are binary: $s_{a,b} \in \{sibling, unrelated\}$. More rarely, ranking of texture retrievals are compared, as this requires human experiments to determine ground-truth and is therefore not often used (see Section 1.1.2).

Hence for the tasks discussed above, binary similarity data are often sufficient and can normally be obtained relatively cheaply compared with, for instance, pair-wise magnitude estimation. However, we believe that the use of such binary data does not provide a very stringent test of a feature set's ability to estimate texture *similarity* and that fine-grained data should be used where possible. This is one of the major reasons as to why we selected *Pertex* [3].

3.2 *Pertex* Textures - Overview

Pertex [3] provides greyscale, 1024×1024 pixel, images of 334 textures captured under constant illumination and viewing conditions. To provide the reader with visual summary of textures contained in the database we performed a simple hierarchical clustering [49]. This used the similarity matrix, 8D-ISO [29], described the next section. Fourteen clusters were obtained by "cutting" the dendrogram at a dissimilarity of 0.34 which in the authors' opinion provided most insight as to the type and range of textures. Fig. 1 shows samples from the resulting clusters together with the dendrogram of the agglomerative clustering. It can be seen that *Pertex* includes a range of semi-structured, unstructured, directional and isotropic textures.

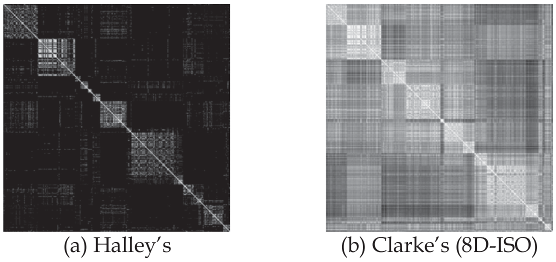
3.3 *Pertex* Ground-Truth

One of the reasons that we selected *Pertex* [3] is because of the availability of two types of fine-grained similarity data. As the two evaluation protocols that we use in Sections 5 and 6 are closely based on these data, we describe both sets in detail below.

3.3.1 Ground-Truth Set 1: Similarity Matrix (8D-ISO)

The first ground-truth set is a matrix of pair-wise similarity estimates, 8D-ISO $\in \mathbb{R}^{334\times334}$, due to the work of Clarke *et al.* [29]. Clarke *et al.* used Isomap [115] to reduce the dimensionality of data derived from an experiment performed by Halley [3], [54] to eight dimensions. In Halley's experiment 30 observers formed groups of textures that they considered "similar". The estimate of the similarity, $s_{a,b} \in \mathbb{Q}$, between textures a and b is defined as the ratio of (1) the number of the observers who have placed *a* and *b* into the same group, to (2) the total number of observers.

Fig. 2 compares plots of (a) Halley's original similarity matrix and (b) Clarke's 8D-ISO similarity matrix. It can be seen that Halley's matrix is sparser, containing many zero entries, which is likely due to the use of only 30 human observers. We chose to use Clarke's 8D-ISO similarity matrix as (1) Pearson's correlation coefficients (calculated between the original similarity matrix and the Isomap versions) increase slowly after eight dimensions [36], (2) reduction to an 8D sub-manifold can be considered to provide a smoothing of inter-observer variation, and (3) because Clarke *et al.* showed


(a) Halley's

(b) Clarke's (8D-ISO)

Fig. 2. Plots of the two 334 x 334 perceptual similarity matrices available for *Pertex* [3]. (a) is as obtained directly from Halley's free-grouping experiment [54], while (b) shows the dimensionality reduced 8D-ISO data later derived by Clarke *et al.* [29]. The brightness of a point denotes the similarity $s_{a,b}$ of two textures and they are ordered by the results of agglomerative clustering in order to make clusters more obvious.

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.

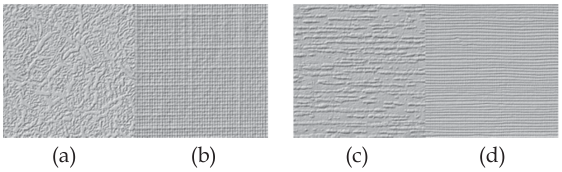


Fig. 3. A pair of pairs set of textures: observers are required to indicate which pair $\{a, b\}$ or $\{c, d\}$ look more similar.

that it agrees well with the second set of ground-truth [29]. Fig. 2b shows that the intra-cluster similarity is still retained while the inter-cluster similarity is not as sparsely represented as before. We attribute the denser similarity matrix to the fact that the Isomap analysis is able to extrapolate the intrinsic relationship between the entities of the similarity matrix when it is applied to the original similarity matrix [54].

3.3.2 Ground-Truth Set 2(a): Human Pair-of-Pairs Judgements ($POPJ_{POP}$)

Clarke *et al.* [29] derived a set of Pair-Of-Pairs Judgements (POPJ) by directly performing a Pair-Of-Pairs (POP) experiment with 20 observers. We use the symbol $POPJ_{POP}$ to refer to these data. In his experiment two pairs of textures $\{\{a, b\}, \{c, d\}\}$ were simultaneously presented to the observer (see Fig. 3). Observers were required to decide which pair, $\{a, b\}$ or $\{c, d\}$, comprised the two textures which they considered to be most similar to each other. Due to the time cost of the experiment (around 2 hours per observer for 1000 trials) and resulting potential observer fatigue, only 1000 pairs-of-pairs (out of approximately 334^4 possible combinations) were used. The 1000 pairs-of-pairs were randomly selected with the restriction that $a \neq b$ and $c \neq d$. Once the experiment was complete, the $POPJ_{POP}$ dataset was directly generated. Entry $POPJ_{POP}(\{a, b\}, \{c, d\})$ was assigned the value of 1 if 11 or more observers thought that pair $\{a, b\}$ was more similar than pair $\{c, d\}$. The value of -1 was assigned for the reverse case, and for ties a value of 0 was assigned.

3.3.3 Ground-Truth Set 2(b): Pair-of-Pairs Data Derived from the 8D-ISO Similarity Matrix ($POPJ_{ISO}$)

We used the human-derived 8D-ISO matrix [29] (see Section 3.3.1) to produce a second source of pair-of-pairs ground-truth. This used the same 1000 pairs-of-pairs textures as Set 2(a) above, but for each pair-of-pairs $\{\{a, b\}, \{c, d\}\}$, we obtained similarities $s_{a,b}$ and $s_{c,d} \in$ 8D-ISO and used them to generate:

$$POPJ_{ISO}\ (\{a, b\}, \{c, d\}) = \begin{cases} 1, & s_{a,b} > s_{c,d} \\ 0, & s_{a,b} = s_{c,d} \\ -1, & s_{a,b} < s_{c,d} \end{cases}. \tag{1}$$

Ground-truth 2(b) both enables comparison with the human pair-of-pairs ground-truth 2(a), and enables assessment of the computational features against 8D-ISO for this subset of data (the 1000 pairs-of-pairs) and in this format.

3.4 Summary

In conclusion, therefore, we used the *Pertex* [3] database due to (1) its coverage of texture types, (2) its use of consistent viewing and image capture conditions, and (3) the availability of its fine-grained human-derived ground-truth. This ground-truth was used in three forms: Set 1, the 8D-ISO similarity matrix derived from a free-grouping experiment [54]; Set 2(a), the pair-of-pairs $POPJ_{\mathrm{POP}}$ comparisons derived directly from human pair-of-pairs judgements [29]; and Set 2(b), $POPJ_{ISO}$, containing information derived from the 8D-ISO [29] coded in a pair-of-pairs format.

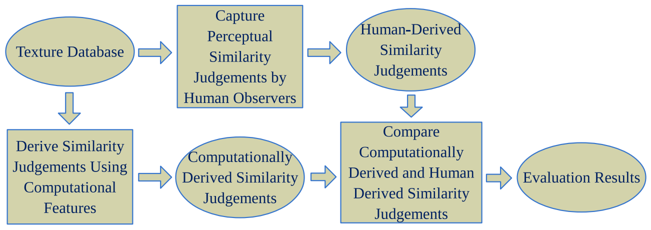

Texture Database
Capture Perceptual Similarity Judgements by Human Observers
Human-Derived Similarity Judgements
Derive Similarity Judgements Using Computational Features
Computationally Derived Similarity Judgements
Compare Computationally Derived and Human Derived Similarity Judgements
Evaluation Results

Fig. 4. The pipeline of the proposed evaluation protocols.

4 Texture Feature Evaluation Protocols

This section describes the two evaluation protocols (see Fig. 4) that we use to assess the texture features reviewed in Section 2. Both use ordinal relationships based on the two ground-truth sets. Set 1 is a similarity matrix. However, rather than compare similarity matrices directly, which can present difficulties for interpretation by users and choices of metric, we chose to compare retrieval *rankings* derived from the similarity matrices. In contrast Set 2 comprises pairs-of-pairs judgements and so for these ground-truth we computed the corresponding results from the computational similarity matrices and compared these directly. In both cases (Sets 1 and 2) the ground-truth could not be computed from pair-wise binary similarity data and therefore we consider them *fine-grained*.

Sections 4.2 and 4.3 describe the two assessment protocols for these two sets of ground-truth. However, first we present the method that we use to generate the multi-resolution Computational Similarity matrices CS that are used in both these protocols.

4.1 Generating Multi-resolution Computational Similarity Matrices

Many of the feature sets listed in Table 1 only compute HOS over small, local regions (as generated in Stage I of the model). It is therefore not uncommon for texture analysis systems to use a multi-resolution scheme designed to allow the features to exploit larger spatial extent. It is also likely that the human visual system processes images using multiple resolutions [67]. Thus, in order to allow a fair assessment of the features, we employ a Gaussian pyramid [20] multi-resolution approach [99]. This generates Computational Similarity (CS) matrices for each of the feature sets $f_i \in \{f_1, f_2 \dots\ f_{51}\}$ at six different image resolutions, $r \in \{1024 \times 1024, 512 \times 512, 256 \times 256,\ 128 \times 128,\ 64 \times 64, Multi\}$ (where $Multi$ refers to all resolutions combined). They are each generated from their corresponding Distance Matrices (DM) as described in Algorithm 1. Each Distance Matrix contains all pair-wise distances ($d_{a,b} \in \mathbb{R}\ \forall\ a, b \in \{t_1, t_2 \dots t_{334}\}$) between all textures t. The *Chi-square* (χ^2) statistic is used to calculate distance values for n-dimensional histogram-based features:

$$d_{a,b} = \frac{1}{2}\sum_{j=1}^{n}\frac{(f_j(\mathrm{a}) - \ f_j(\mathrm{b}))^2}{f_j(\mathrm{a}) + f_j(\mathrm{b})}. \tag{2}$$

While the *Euclidean* distance is used for other feature sets:

$$d_{a,b} = \sqrt[2]{\sum\nolimits_{j=1}^{n}(f_j(a) - f_j(b))^2}. \tag{3}$$

The resulting 6 × 51 Computational Similarity matrices $CS \in \mathbb{R}^{334\times334}$ are used in both Protocols I and II.

4.2 Protocol I: Texture Retrieval Based Evaluation

Inspired by the use of human ordinal data for the evaluation of the performance of search engines [13], [58], [87], we employ a texture retrieval based evaluation protocol. This allows us to compare the use of computational and human-derived similarity matrices in an applicable task: the retrieval of the top N textures in response to the presentation of one of the 334 textures as a *query* image. Note that these rankings could not be produced unambiguously if binary pair-wise similarities were used.

The top N textures are derived from a similarity matrix ($S \in \mathbb{R}^{334\times334}$) by extracting the row (or column) containing the *query* image q, and ordering the remaining 333 textures according to their similarities to q. We compared all 334 rankings obtained from the human-derived 8D-ISO similarity matrix [29], against the corresponding rankings obtained from each of the 6 × 51 Computational Similarity matrices, for $N \in \{10,\ 20,\ 40,\ 60\}$. The comparisons were performed using the two measures described below.

4.2.1 Performance Measures: G and M

We chose to use the performance measures G [43] and M [13] to assess ranking performance. These measures can compare two rankings which contain *different* elements and consider not only the number of the relevant items retrieved, but also the ranking *positions* of these elements. Thus, we believe that they provide a more informed measure than more commonly used metrics, such as Precision [8], Recall [8], Normalised Precision [79], and Normalised Recall [79]. The G measure [43] is defined as:

$$G \ = \ 1 - \ \frac{\sum_{i\,=\,1}^{R}(|R_C(i) - R_{ISO}(i)|) + \sum_{i\,=\,1}^{N-R}[(N+1) - R_C(i)] + \sum_{i\,=\,1}^{N-R}[(N+1) - R_{ISO}(i)]}{N(N+1)}, \tag{4}$$

where R is the number of relevant images in N retrieved images, $R_C(i)$ is the rank of the *i*-th image retrieved by a computational feature set, and $R_{ISO}(i)$ is the rank of the *i*-th image retrieved using the 8D-ISO similarity matrix [29].

The M measure [13] was motivated by the observation that identical (exactly same in elements and ranks) or nearly-identical rankings of the top N images are more important to humans than those among lower placed images. Using the same notation as in Equation (4), M is defined as:

$$M \ = \ 1 - \ \frac{\sum_{i\,=\,1}^{R}\left(\left|\frac{1}{R_C(i)} - \frac{1}{R_{ISO}(i)}\right|\right) + \sum_{i\,=\,1}^{N-R}\left(\frac{1}{R_C(i)} - \frac{1}{N+1}\right) + \sum_{i\,=\,1}^{N-R}\left(\frac{1}{R_{ISO}(i)} - \frac{1}{N+1}\right)}{2\sum_{i\,=\,1}^{N}(\frac{1}{i} - \frac{1}{N+1})}. \tag{5}$$

Algorithm 1. The Algorithm for Computing Similarity Matrices

```
(1) Each texture image is decomposed into five Gaussian pyramid sub-bands corresponding to resolutions $1024 \times 1024$, $512 \times 512$, $256 \times 256$, $128 \times 128$ and $64 \times 64$;
(2) Each sub-band is individually normalised to have an average intensity of 0 and standard deviation of 1 in order to remove the influence of 1st- and 2nd-order grey level (moment) statistics;
(3) Feature extraction is performed to obtain a feature vector from each sub-band separately, and all five feature vectors are also combined into an additional "multi-resolution" feature vector. Thus, in total six feature vectors are derived for each texture;
(4) A pair-wise distance matrix $DM$ is computed from all 334 sub-band images at each pyramid level or in the multi-resolution case. Each $DM$ is normalised to the range of [0, 1] and is then converted into a similarity matrix $SM$ according to $SM = 1.0 - DM$. Hence, six computational similarity matrices are obtained for each feature set.
```

Algorithm 2. The algorithm for texture retrieval evaluation

```
For each computational feature set, $f \in \{51 \text{ feature sets}\}$; each resolution $r \in \{1024 \times 1024,\ 512 \times 512,\ 256 \times 256,\ 128 \times 128,\ 64 \times 64, Multi\}$; and each retrieval set size, $N \in \{10,\ 20,\ 40,\ 60\}$; do:
(i) For each query texture, $q \in \{334\ Pertex\ \text{textures}\}$ and with $q$ excluded from the retrieval set;
    (a) Use the 8D-ISO similarity matrix to obtain the ranked list ($R_{ISO}$) of the first N textures;
    (b) Use the computational similarity matrix $CS_{f,r}$, to derive the ranked list ($R_C$) of the top $N$ textures;
    (c) Compute the $G$ and $M$ scores for comparing the two ranked lists: $R_{ISO}$ and $R_C$;
(ii) Average the G and M scores over all 334 query textures and use these averages as performance metrics.
```

4.2.2 Comparison Process

The evaluation process is conducted as described in Algorithm 2. Thus, for the retrieval evaluation described in Section 5 it produces 6 × 51 × 4 pairs performance measures, $\{G, M\}$. That is it produces performance measures at each of six resolutions, for each of 51 feature sets and for each of four values of N.

4.3 Protocol II: Pair-of-Pairs Based Evaluation

The second type of ground-truth (Set 2) comprises two sets of pairs-of-pairs judgements ($POPJ_{POP}$ and $POPJ_{ISO}$). As it

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.

TABLE 3
Best Feature Sets for Texture Retrieval

N	Measure	Resolution					
		$\mathbf{1024 \times 1024}$	$\mathbf{512 \times 512}$	$\mathbf{256 \times 256}$	$\mathbf{128 \times 128}$	$\mathbf{64 \times 64}$	Multi
10	G	VZ-NBRHD 0.21	VZ-MRF 0.21	VZ-MRF 0.20	LBPBASIC 0.20	LBPBASIC 0.16	LBPHF 0.23
	M	VZ-NBRHD 0.19	VZ-MRF 0.20	VZ-MRF 0.19	LBPBASIC 0.18	LBPBASIC 0.13	LBPBASIC 0.20
20	G	VZ-NBRHD 0.25	VZ-MRF 0.25	MRSAR 0.24	LBPBASIC 0.24	LBPBASIC 0.20	MRSAR 0.28
	M	VZ-NBRHD 0.20	VZ-MRF 0.21	VZ-MRF 0.20	LBPBASIC 0.20	LBPBASIC 0.15	LBPHF 0.22
40	G	VZ-NBRHD 0.30	VZ-MRF 0.30	MRSAR 0.32	MRSAR 0.32	MRSAR 0.28	MRSAR 0.36
	M	VZ-NBRHD 0.22	VZ-MRF 0.23	VZ-MRF 0.22	LBPBASIC 0.22	LBPBASIC 0.18	MRSAR 0.25
60	G	RING & WEDGE 0.35	RFS 0.36	MRSAR 0.38	MRSAR 0.38	MRSAR 0.34	**MRSAR** **0.41**
	M	VZ-NBRHD 0.24	VZ-MRF 0.25	MRSAR 0.24	LBPBASIC 0.24	MRSAR 0.20	**MRSAR** **0.27**

This table shows the best performing feature sets for Protocol I applied to assessing the ability of 51 Computational Similarity matrices to retrieve N textures (from 334) for 4 values of N, and 6 resolutions, r (here "Multi" means multi-resolution). Bold digits indicate the best G or M scores.

is easy to derive equivalent data from Computational Similarity (CS) matrices we chose to examine the *agreement rate* between these data and the human-derived ground-truth.

4.3.1 Generating Pair-of-Pairs Judgements ($POPJ_C$) from Computational Similarity Matrices (CS)

We generated 1000 pairs-of-pairs $\{\{a, b\}, \{c, d\}\}$ judgements (corresponding directly to the 1000 trials performed in the human $POPJ_{POP}$ experiment [29]) from each of the 6 × 51 Computational Similarity matrices. That is, for each Computational Similarity matrix (CS), we extracted the similarities $s_{a,b}$ and $s_{c,d} \in CS$, corresponding to each of the 1000 $\{\{a, b\}, \{c, d\}\}$ trials. These were used to generate 6 × 51 sets of "computational pairs-of-pairs judgements" ($POPJ_C$) using Equation (1). That is for each of the 51 feature sets listed in Table 1, a total six sets of $POPJ_C$ are obtained (one for each of the five single pyramid resolutions and one for the multi-resolution scheme).

4.3.2 Comparing Computational and Human-Derived Pair-of-Pairs Judgements

We use *Agreement Rate* (AR) to refer to the normalised agreement, over 1000 trials, between computational and human-derived pair-of-pairs judgements. Thus for each feature set, at each resolution, and for each of the two ground-truth sets $POPJ_{POP}$ and $POPJ_{ISO}$ we compute AR:

1) Compute agreement between computational and human-derived pair-of-pairs judgements at trial *i*:

$$IsAgreed(i) \ = (POPJ_H \ (i) \ == POPJ_C \ (i))?1 : 0, \qquad (6)$$

where: $POPJ_H \in \{POPJ_{POP}, POPJ_{ISO}\}$;

2) Calculate the agreement rate (AR):

$$AR \ = \frac{\sum_{i=1}^{1000} IsAgreed(i)}{1000} \ . \qquad (7)$$

In this paper, we use a percentage (%) to denote agreement rates.

4.4 Comparing Protocols I and II

We note that for Protocol I (retrieval) only the top N most similar textures are retrieved, and it is therefore likely that this approach tests intra-cluster similarities more than inter-cluster similarities (these clusters are shown in Fig. 1). However, the *pair-of-pairs* evaluation protocol examines both intra-cluster and inter-cluster similarities.

5 Results of Protocol I: Texture Retrieval Based Evaluation

In this experiment we use Protocol I to compare the performance of the computational feature sets with human-derived ground-truth to discover whether or not certain feature sets or feature categories generally perform better than their counterparts and to determine what the effect of resolution has.

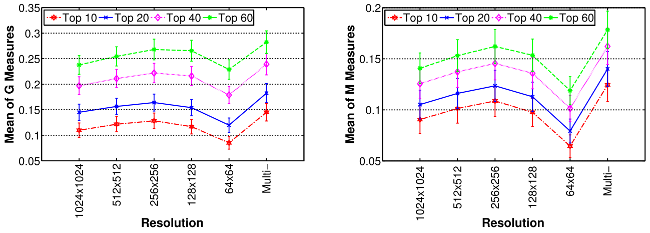
5.1 Highest Scoring Feature Sets

Table 3 reports the highest G and M scores achieved by the 51 feature sets, for each of four different retrieval set sizes and at each of six different resolutions, when the ground-truth 8D-ISO [29] is used. From this it can be seen that there is no single "best" feature for all resolutions. However, perhaps the most significant result is that even for the case providing the highest G and M scores (of 0.41 and 0.27 for MRSAR [83]) the average proportion of the number of the relevant textures retrieved was only 48 percent.

5.2 Effect of the Image Resolution

We examine the effect of the resolution on texture retrieval performance. We first test the significance of the effect on the G and M scores, and then empirically investigate the effect of the multi-resolution scheme.

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.



(a) Mean G scores (b) Mean M scores

Fig. 5. Means and 97.5 percent confidence bounds of the G and M scores calculated over the 51 feature sets, at each resolution and for each retrieval size, $N \in \{10,\ 20,\ 40,\ 60\}$. These figures show the superiority of the multi-resolution approach.

5.2.1 The Significance of Resolution

Since we were not interested in the interactions between the G and M scores, we performed two factorial repeated-measures ANOVAs [45] on the G and M scores separately. The family-wise error needs to be controlled by adjusting the level of significance for each ANOVA to ensure that the overall Type I error rate (∂) throughout both ANOVAs stays at 0.05. Hence, $\partial =\ 0.025$ was used as the Bonferroni correction [45]. The results of Mauchly's test [45], [85] applied to the G scores show that the assumption of sphericity is violated for the main effect of the resolution, $\chi^2(14)\ =\ 186.38, P = 0.00$. Degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity [53] ($\varepsilon =\ 0.46$). The results show that resolution has a significant effect on G scores, $F(2.28, 114.03) = 41.87, P = 0.00$. The Mauchly's test performed on M scores shows that the assumption of sphericity was also violated $\chi^2(14)\ =\ 169.54, P = 0.00$. Degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon =\ 0.45$). A significant main effect of resolution on M scores is also observed, $F(2.25, 112.46) = 51.77, P = 0.00$. In both cases ($G$ and M) contrasts [45] reveal that the multi-resolution scheme performs better than any of the individual resolutions. The superiority of the multi-resolution approach is also apparent in Fig. 5 which shows the 97.5 percent confidence bounds of both G and M scores.

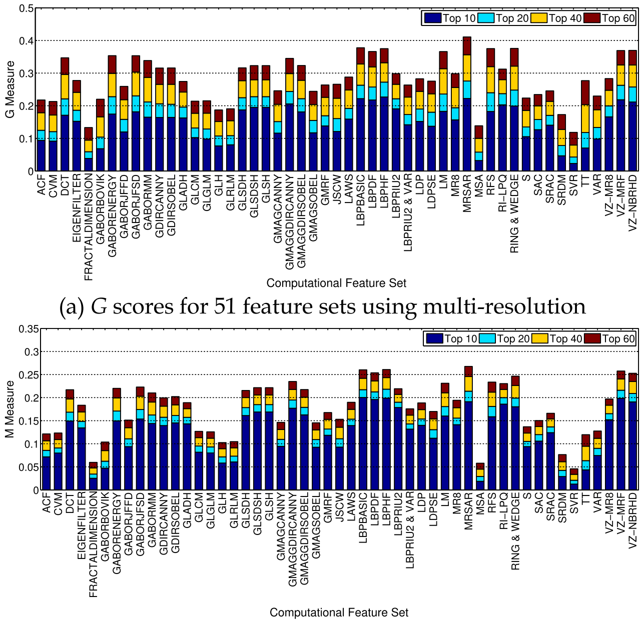
5.2.2 Improvement Over Original Resolution

In a further investigation of the effect of resolution, we examined the individual G and M scores obtained using the 51 feature sets with the original 1024×1024 images, against their performances at the five other resolutions $r \in \{64 \times 64,\ 128 \times 128,\ 256 \times 256,\ 512 \times 512,\ Multi\}$. Table 4 reports the numbers of feature sets that were superior at each of these resolutions compared with the original 1024×1024. It shows that the performances of at least 46 (out of the 51) feature sets were boosted by using the multi-resolution scheme and that this improvement was greater than for each of the other four resolutions.

TABLE 4
Numbers of Feature Sets Whose G or M Scores Were Enhanced Using the Resolutions $r \in \{64{\times}64,\ 128{\times}128,\ 256{\times}256,\ 512{\times}512,\ Multi\}$ Compared With the Scores Obtained Using the Original Resolution of $1024{\times}1024$

N	Measure	Resolution				
		512×512	256×256	128×128	64×64	Multi
10	G	42	40	30	14	**47**
	M	38	44	29	11	**46**
20	G	42	40	31	14	**47**
	M	41	43	31	12	**46**
40	G	44	41	34	20	**48**
	M	44	43	32	13	**46**
60	G	45	44	38	23	**49**
	M	44	43	32	16	**46**

Bold digits indicate the largest numbers of the enhanced feature sets.



(a) G scores for 51 feature sets using multi-resolution

(b) M scores for 51 feature sets using multi-resolution

Fig. 6. Multi-series bar chart of the G and M scores obtained using Protocol I with 51 feature sets and $N \in \{10,\ 20,\ 40,\ 60\}$. Note that only the multi-resolution approach ($r = Multi$) was used here. These results show that even at the best resolution, none of the feature sets perform as well as might be expected.

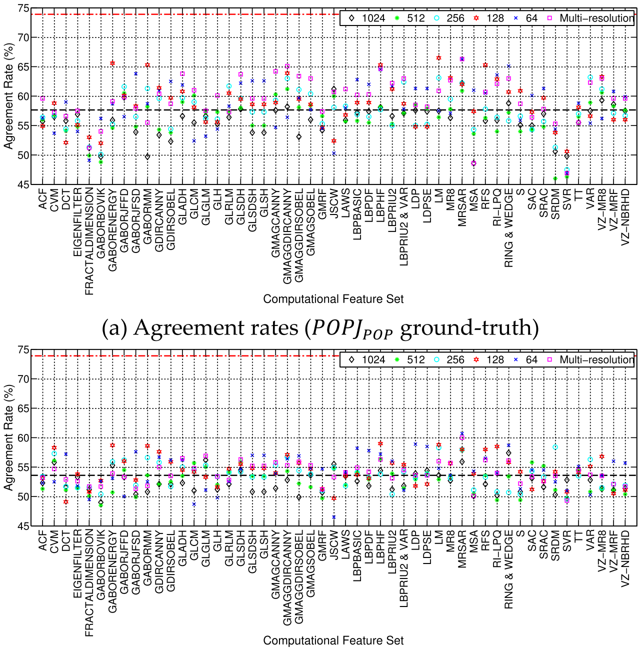
5.2.3 Comparing Computational Features Using the Multi-Resolution Scheme

Given that the multi-resolution approach provided the best retrieval results, we decided to investigate the associated individual performances of the feature sets. Fig. 6 shows the G and M scores derived for $r = Multi$ for all 51 feature sets. This again shows MRSAR [83] to be the highest scoring feature set when 60 textures were retrieved, but in addition shows that there is considerable variation between features.

5.3 Summary of the Texture Retrieval Evaluation

From the above it can be concluded that the multi-resolution, pyramid, approach clearly gives the best results when employing Protocol I. However, even when only considering the best performances obtained using the multi-resolution approach, (1) the orders of textures in computational and perceptual rankings differ considerably, and (2) no more than one half of retrieved textures are relevant. In other words, even the performance of the best feature set (MRSAR [83]) using multi-resolution is still considerably below what might be expected from texture features that often reported to achieve over 90 percent success rates in classification [83], [91], [122], [123] and segmentation [61] tasks.

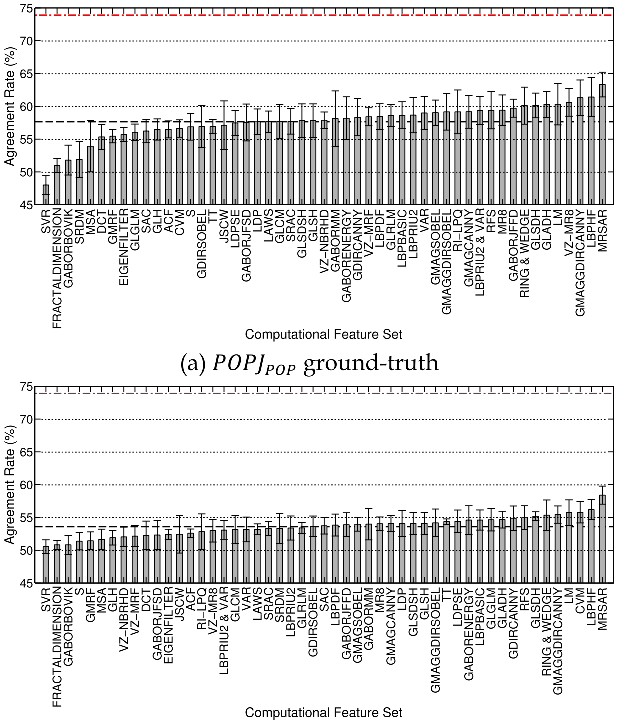
Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.



(a) Agreement rates ($POPJ_{POP}$ ground-truth)

(b) Agreement rates ($POPJ_{ISO}$ ground-truth)

Fig. 7. Agreement rates (%) for the two different types of ground-truth obtained using 51 feature sets and at six different "resolutions". The black bold dashed lines show average agreement rates: (a) 57.7 percent and (b) 53.6 percent (computed across 51 feature sets and six resolutions). For reference, the top red bold dash-dot line in each graph shows the agreement between the two ground-truth data sets (73.9 percent).



(a) $POPJ_{POP}$ ground-truth

(b) $POPJ_{ISO}$ ground-truth

Fig. 8. Average agreement rates (%) for 51 feature sets (sorted in an ascending order) and their 95 percent confidence bounds, for six resolutions. The black bold dashed lines show the average agreement rates 57.7 and 53.6 percent (calculated over the 51 feature sets and six resolutions. For reference, the top red bold dash-dot line in each graph shows the agreement between the two ground-truth data sets (73.9 percent).

6 Results of Protocol II: Pair-of-Pairs Based Evaluation

In this section we use Protocol II to assess the capabilities of the computational feature sets to estimate two sets of pair-of-pairs ground-truth: $POPJ_{POP}$ (obtained directly from experiment [29]) and $POPJ_{ISO}$ (obtained by free-grouping [54] followed by Isomap [115] analysis). Specifically, we investigate: (1) whether any of the feature sets achieve performances near ground-truth; (2) whether there is a universally "best" feature set or feature category that works across resolutions; and (3) the optimal resolution. Finally, we investigate how the feature sets perform in combination by using Random Forest regressors [18], and additionally investigate the performance of two image quality assessment measures.

6.1 Best Performance of Feature Sets

Fig. 7a shows the agreement rates (%) obtained using the $POPJ_{POP}$ judgements. The highest agreement rate 66.5 percent was obtained using LM [72] at the resolution of 128×128 pixels, while the lowest agreement rate 46.0 percent was derived using SRDM [66] at the resolution of 512×512 pixels. Fig. 7b shows agreement rates (%) for the $POPJ_{ISO}$ ground-truth. It can be seen that the highest agreement rate, 60.7 percent, was obtained using MRSAR [83] while the lowest agreement rate of 46.5 percent was derived using JSCW [99] at the resolution of 64×64 pixels.

Compared with agreement between the two sets of human judgements (red bold dash-dot line), the agreement rates of the 51 feature sets vary over a relatively small range. This result indicates that common properties may be exploited by these feature sets, e.g., the calculation of local region HOS as discussed in Section 2.5. But as the performance of *each* feature set varies significantly within this range we cannot reliably determine the "best" feature set that reliably works *across* the six resolutions.

Moreover, the average agreement rates obtained for the two different types of ground-truth across all feature sets and resolutions are 57.7 percent ($POPJ_{POP}$) and 53.6 percent ($POPJ_{ISO}$). This is considerably below the 73.9 percent agreement rate between the two ground-truth sets. (We attribute differences between ground-truth to either use of different observer groups or method).

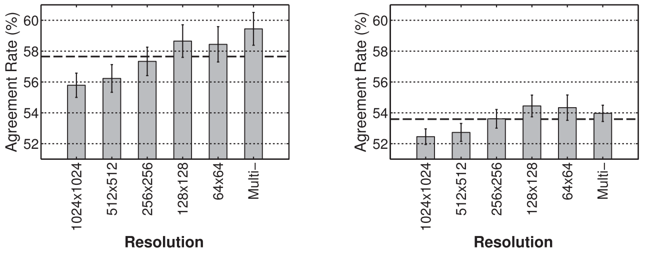
Thus, the most obvious observation from this analysis is that the performance of the feature sets differs considerably from that of human observers.

6.2 Average Performance across Resolutions

In order to remove the effect of the resolution, average agreement rates and 95 percent confidence bounds were computed across six resolutions for each feature set (Fig. 8). When $POPJ_{POP}$ judgements were used (Fig. 8a), MRSAR [83] outperformed its counterparts with an agreement rate of 63.3 percent while the worst performance, 48.0 percent, was obtained using SVR [90]. Similarly, when $POPJ_{ISO}$ judgements were used (see Fig. 8b), MRSAR [83] obtained the highest agreement rate, 58.4 percent, and SVR [90] was outperformed by its counterparts at an agreement rate of 50.6 percent.

However, while these results indicate that MRSAR *may* be the best performing feature, its 95 percent bound overlaps with those of the following feature sets. Hence, again we cannot treat it as being reliably the best feature set overall.

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.



(a) $POPJ_{POP}$ ground-truth (b) $POPJ_{ISO}$ ground-truth

Fig. 9. The effect of resolution: average agreement rates and 95 percent confidence bounds are shown for each of six resolutions computed over the 51 feature sets. The black bold dashed lines indicate overall averages: 57.7 percent (a) and 53.6 percent (b).

TABLE 5
Numbers of Feature Sets Whose Agreement Rate Was Enhanced Using the Resolutions $r \in$ {64×64, 128×128, 256×256, 512×512, *Multi*} Compared With That Obtained Using the Original Resolution of 1024×1024, Together With Two Sets of Ground-Truth

Resolution	512×512	256×256	128×128	64×64	Multi
$POPJ_{POP}$	31	34	40	38	**49**
$POPJ_{ISO}$	27	37	39	39	**43**

6.3 Effect of Image Resolution

Fig. 9 shows agreement rates aggregated over the resolutions $r \in \{1024 \times 1024,\ 512 \times 512,\ 256 \times 256,\ 128 \times 128,\ 64 \times 64,\ Multi\}$ for both sets of ground-truth ($POPJ_{POP}$ and $POPJ_{ISO}$). In both cases the original resolution $r = 1024 \times 1024$ is outperformed by resolutions that allow greater spatial extent to be exploited by Stage I processing.

To test the significance of this effect, we performed two one-way repeated-measures ANOVAs [45]. Mauchly's test [45], [85], indicated that the assumption of sphericity was violated ($\chi^2(14) = 78.63$ and $\chi^2(14) = 64.39, P = 0.00$) when $POPJ_{POP}$ and $POPJ_{ISO}$ were used as ground-truth respectively. Degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.62$ and 0.65 respectively). The results show that agreement rates were significantly affected by resolution in both cases: $F(3.11, 155.71) = 15.97$ and $F(3.25, 162.66) = 9.61, P = 0.00$.

The post hoc tests performed using the Bonferroni correction reveal that the agreement rates derived at $r \in \{256\times 256,\ 128 \times 128,\ 64 \times 64, Multi\}$ were significantly different from those obtained at the original 1024×1024 resolution, $P < 0.05$. However, there was no significant difference detected between the four resolutions, $P = 1.00$. These results held for both sets of ground-truth. Table 5 shows the number of feature sets whose pair-of-pairs predictions were improved.

In conclusion, therefore, as the significance tests for both pair-of-pairs ground-truth indicate that we should use one of $r \in \{256 \times 256,\ 128 \times 128,\ 64 \times 64, Multi\}$, and as multi-resolution proved significantly better in the retrieval evaluations (Protocol I), we believe that the multi-resolution approach should be investigated before others.

6.4 Comparing Results Obtained: Set 2(a) Versus 2(b)

Table 6 shows the results testing the correlations between agreement rates obtained using ground-truth Set 2(a), $POPJ_{POP}$ against those obtained using Set 2(b), $POPJ_{ISO}$ (for six resolutions and the means across these resolutions). It shows that the results obtained for the 51 feature sets are closely correlated ($P < 0.05$) with each other across the two ground-truth sets at all resolutions. It also indicates that the two sets of human perceptual data are consistent with each other.

TABLE 6
Spearman's Correlation Coefficients (ρ, $\partial = 0.05$) and *P* Values: (Columns 2-7) Between the Two Sets of Curves in Fig. 7a and 7b; and (column 8) Between the Two Curves in Fig. 8a and 8b

	1024×1024	512×512	256×256	128×128	64×64	Multi	Mean
ρ	0.58	0.49	0.56	0.85	0.82	0.70	0.61
P-Val	0.00	0.00	0.00	0.00	0.00	0.00	0.00

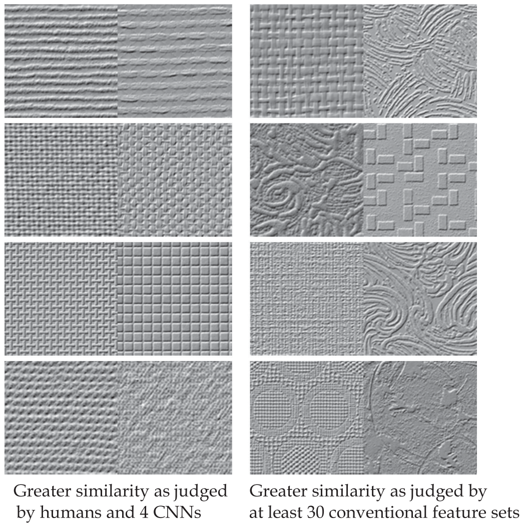

Greater similarity as judged by humans and 4 CNNs
Greater similarity as judged by at least 30 conventional feature sets

Fig. 10. The top four worst pair-of-pairs results in which the majority of human observers considered that the left pairs are more similar while at least 30 of the 51 conventional feature sets did not. Note that only the central quarters of textures are shown. (Also note that the CNN results are discussed in Section 8).

6.5 Worst Human and Feature Disagreements

To give some insight as to the worst errors of the feature sets, we show the four pairs-of-pairs that caused the "worst" disagreements in Fig. 10. We determined the "worst" pairs-of-pairs by selecting ones in which (1) the disagreements between at least 30 of the 51 feature sets disagreed with the majority human decision; and (2) the observers in the free-grouping experiment [54] agreed with the majority decision in the pair-of-pairs experiment [29]. These were then sorted in descending order of the average disagreement between humans and features. (For more details, see Appendix E in [36]).

6.6 Test of Image Quality Assessment Measures

Additionally, at this stage we tested two classical image quality measures: SSIM [126] and MS-SSIM [127] as they are designed to assess the similarity of two images $\{a, b\}$. Their outputs were used directly to calculate $s_{a,b}$. Note that for MS-SSIM we used the default parameters including five scales [127]. Table 7 reports the agreement rates derived using these two measures. As can be seen, these results are inferior to the

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.

TABLE 7
Agreement Rates (%) Obtained Using the SSIM [126] and MS-SSIM [127] Quality Measures

Measure	SSIM					MS-SSIM
Resolution	$\mathbf{1024 \times 1024}$	$\mathbf{512 \times 512}$	$\mathbf{256 \times 256}$	$\mathbf{128 \times 128}$	$\mathbf{64 \times 64}$	Multi
$POPJ_{POP}$	56.6	**60.8**	60.7	49.7	53.1	40.4
$POPJ_{ISO}$	53.8	57.3	**58.1**	49.6	53.1	44.1

TABLE 8
Average Agreement Rates (%) Obtained Using Random Forests With 20, 50, 100, or 200 Trees Compared With the Human-Derived Ground-Truth: $POPJ_{POP}$

Num. of Trees	$\mathbf{1024 \times 1024}$	$\mathbf{512 \times 512}$	$\mathbf{256 \times 256}$	$\mathbf{128 \times 128}$	$\mathbf{64 \times 64}$	Multi
20	66.6 ± 0.8	66.6 ± 1.5	$\mathbf{66.9 \pm 2.2}$	66.3 ± 1.6	61.0 ± 1.3	66.3 ± 2.4
50	67.3 ± 1.6	$\mathbf{67.8 \pm 2.4}$	66.8 ± 1.7	66.2 ± 1.6	62.9 ± 1.5	66.8 ± 2.7
100	65.9 ± 2.0	$\mathbf{67.8 \pm 2.5}$	66.5 ± 2.7	66.2 ± 1.5	63.4 ± 1.6	67.7 ± 1.3
200	66.8 ± 1.9	67.0 ± 1.9	$\mathbf{67.7 \pm 2.1}$	65.9 ± 0.8	63.1 ± 2.3	67.2 ± 1.9

agreement rate of 73.9 percent computed between the two pair-of-pairs ground-truth sets.

6.7 Random Forest Regression

The use of Protocols I and II showed that, individually, none of the feature sets achieved agreement rates similar to that obtained between the two human-derived ground-truth sets. Thus it was decided to investigate the performance of the union of these feature sets $\boldsymbol{f} = \{f_1 \cup f_2 \ldots f_{51}\}$ using Random Forest regression [18].

We used three-fold cross-validation to train the regressors to predict similarity values ($\hat{s}_{a,b}$) given normalised feature vectors $\boldsymbol{f}'(a)$ and $\boldsymbol{f}'(b)$. Protocol II was used to assess the resultant similarity values.

Feature vectors were pre-processed using two-step normalisation [28]. For each texture image (t) each feature vector, $f_i(t) \in \{f_1(t), f_2(t) \ldots\ f_{51}(t)\}$, was L_2 normalised. These were concatenated into a single feature vector $\boldsymbol{f}(t) = \{f_1(t) \cup f_2(t) \ldots f_{51}(t)\}$ which was further L_2 normalised to provide a final feature vector $\boldsymbol{f}'(t)$ for each of the 334 *Pertex* [3] textures.[2]

Training and test data were produced by randomly partitioning *Pertex* [3] into three subsets of *A*, *B* and *C*, comprising 111, 111 and 112 textures respectively, and creating three corresponding submatrices of 8D-ISO $\in \mathbb{R}^{334 \times 334}$ [29]. Restricted by the hardware, however, only one submatrix (A, B or C) was used in each fold for training while the remaining two were used for testing.

The regressors resulting from each fold were used to predict the similarities $\hat{s}_{a,b}$ and $\hat{s}_{c,d}$ for all the texture pairs $\{\{a,b\},\{c,d\}\}$ contained in ground-truth $POPJ_{POP}$. Agreement rates were calculated as described in Protocol II using this ground-truth. This was performed using 20, 50, 100 and 200 trees, and for six resolutions: $r \in \{1024 \times 1024,\ 512 \times 512, 256 \times 256, 128 \times 128,\ 64 \times 64, Multi\}$.

Table 8 shows the agreement rates of the Random Forest regressors [18]. The best mean performance (across the three folds) is $67.8 \pm 2.4\%$. This performance is superior to the best result, 66.5 percent, produced by the 51 feature sets, but still lower than the agreement rate (73.9 percent) of the two ground-truth sets ($POPJ_{POP}$ and $POPJ_{ISO}$).

Our conclusion, therefore, is that while employing a wide variety of texture features may be advantageous, it does not guarantee optimal performance.

6.8 Summary of the Pair-of-Pairs Evaluation

This section has reported the results obtained using Protocol II with two sets of ground-truth each containing 1000 pairs-of-pairs judgements. They suggest (1) that the two data sets are consistent, (2) that the best results using individual feature sets are achieved using a multi-resolution scheme, and (3) that even employing large numbers of feature types using a Random Forest does not guarantee optimal similarity estimation.

7 Investigating the Importance of Long-Range Interactions to Texture Similarity

The analysis presented in Section 2 (Table 1) shows that the majority of the 51 feature sets examined do not encode long-range aperiodic image structure. In contrast humans are thought to be able to utilise long-range interactions [46], [98], [113]. We therefore hypothesise that the poor performance of the feature sets reported above, may be due to features not using this type of data, even when a multi-resolution pyramid is employed.

Classical receptive-field models are thought to account for local perceptual effects; however, they are not suited for explaining global effects [113]. Similarly, the relationships between the perception and centre-surround antagonism of retinal receptive fields are thought to be limited to sensing short-range interactions. However, it is known that humans are able to utilise long-range interactions for other tasks [46], [98], [113]. Furthermore, such interactions have been used as geometric constraints to guide texture synthesis [14], [73], [74], [76], [78]. In particular, it has been shown that these data are important to perceptual texture synthesis quality assessment [73]. Recently, Kohler *et al.* [68] found that the patch-based feature set [123] cannot represent the parametric

[2] Although the normalisation of feature vectors is not necessary for random forests as they do not compute the distance between feature vectors, we conducted this operation for the fair comparison with other classifiers (see supplemental material).

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.

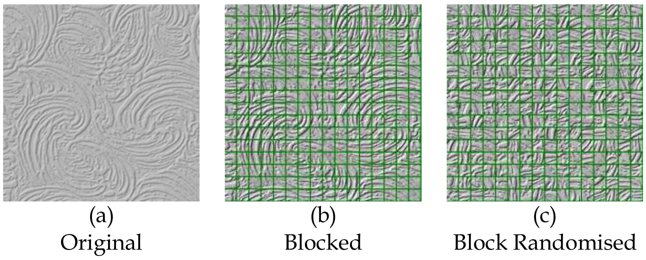


Fig. 11. An original *Pertex* [3] image (a) shown "Blocked" in (b) and then "Block Randomised" in (c). Note that the long-range interactions are easily perceived in (a) and (b) but not (c).

dependence of the responses to rotation symmetry. This representation requires a series of higher-order ventral stream areas, V4, VO1, and lateral occipital complex (LOC).

The above implies that humans do use longer-range interactions and that this information may be important in the judgement of texture similarity. We therefore conducted two studies to investigate this further. Study I was designed to understand if humans use long-range interactions in texture similarity judgement (pair-of-pairs comparison [29]) tasks. Study II investigated whether or not agreement rates, between feature sets and humans, increase when long-range interactions are removed from (or at least reduced in) texture images.

7.1 Block Randomisation

The obvious approach for both studies is to design stimuli (textured images) that contain only short-range interactions, and then to add long-range interactions in a controlled manner. Unfortunately this is difficult, if not impossible to do, affecting local characteristics and thereby introducing confounding factors. However, we can remove, or at least scramble, long-range interactions in existing textures by partitioning the image into small blocks and then randomising their positions. Unfortunately the boundary between two randomised blocks introduces new short-range interactions which could affect human perception. This is likely to be the case even if we use texture synthesis [40] to make the "boundary area" change gradually. Thus, in order to remove (or at least reduce) long-range interactions while inhibiting perceived *changes* in short-range interactions, we first overlay the texture image with a green grid and then randomise the position of blocks as shown in Fig. 11. Note that Field *et al.* [46] used the concept of the "association field" to explain how continuity may be encoded by a visual system and showed that humans can still recognise the pattern in an image even if a grid has been imposed on top of it. Thus, we believe that "Block Randomised" images provide a way of creating effective experimental stimuli with and without long-range interactions.

7.2 Study I

This study investigated whether or not reducing long-range interactions, using Block Randomisation affects humans' pair-of-pairs judgements. We performed the study in two sessions: Session I used Block Randomised images (Fig. 11c) while, as a control, Session II used purely Blocked images (Fig. 11b). For this we selected the 80 pairs-of-pairs texture combinations (out of 1000 $POPJ_{POP}$ trials [29]) that had produced the "worst" disagreements between computational features and humans. (See Section 6.5 for a definition of "worst" in this context).

7.2.1 Study I: Design

Stimuli. All original texture images were blocked with a green grid. The reasons for using green rather than other psychological primary colours are that (1) it is more comfortable [7] and hence impairs human perception less; and (2) it makes the grid easy to discriminate from the grey texture. The thickness of the grid was three pixels and the size of the blocks was 19×19 pixels which is the largest spatial extent used by the majority of the 51 feature sets for computing higher order statistics (see Table 1).

Observers. Ten observers with normal or corrected-to-normal vision were used. None of these observers had attended the original pair-of-pairs experiment [29]. All observers signed a consent form before they performed the experiments. Each observer was given a 5 GBP Amazon voucher after they completed.

Procedure. Session I (using Block Randomised images) was conducted at least one week earlier than the control (Session II which used Blocked images) in order to reduce learning effects. The 80 trials were shown in random order to each observer in each stage. Throughout all 80 trials, observers were simultaneously presented two texture image pairs (left and right) and were then required to decide which pair was more similar.

Tools. All stimuli were shown on a calibrated NEC LCD2090UXi monitor at the resolution of 512×512 pixels. The monitor has a resolution of 1600×1200 pixels and pixel dimensions are 0.255 mm$\times$0.255 mm (100 dpi). Thus, the size of all stimuli was 130.56 mm$\times$130.56 mm when shown on the monitor. Besides, the monitor was linearly calibrated with gamma $= 1$ by a Gretag-MacBeth Eye-One, with a maximum luminance of $120\ \mathrm{cd/m^2}$.

Environment. The distance between observers and the monitor was approximately 50 cm, providing an angular resolution of around 17 cycles per degree. Hence, the stimuli subtended an angle of 14.89° in the vertical direction. The eyes of the observers were located approximately along the axis of the centre of the screen. Both experiments were conducted in a dark room with opaque, black curtains and matte walls.

7.2.2 Study I: Results

For Session I, Block Randomised, agreement rates, $AR_R(ob)$, for each observer $ob \in \{1, 2, \ldots 10\}$, where calculated using Equation (7) and the relevant $POPJ_{POP}$ data [29]. The agreement rates $AR_B(ob)$ for Session II (Blocked) were calculated in the same manner. The means and 95 percent confidence bounds of AR_R and AR_B are shown in Fig. 12.

7.2.3 Study I: Analysis

The K-S test [69], [110] was used to test the normality of the agreement rates: AR_B and AR_R. In addition, this test was applied to the difference between AR_B and AR_R as the *t*-test between AR_B and AR_R was dependent. The results are reported in Table 9. It indicates that the three sets of data follow the normal distribution.

A dependent *t*-test [62] ($\partial = 0.05$) was performed on the two sets of data: AR_B and AR_R. The result shows a

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.

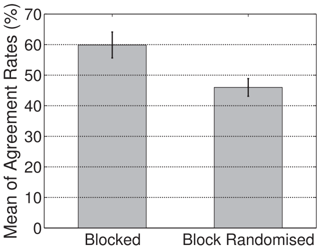


Fig. 12. Study I results: Means and 95 percent confidence bounds of the agreement rates: AR_B (left) and AR_R (right) compared with the original results reported in Section 6.

significantly higher agreement (with the original pair-of-pairs experiment [29]) when non-randomised, but Blocked images, were used ($M = 59.88, SE = 2.17$) compared with the use of Block Randomised images ($M = 46.00, SE = 1.48$), $t(9) = 12.008, P = 0.000, r = 0.970$.

This finding, together with Fig. 12, suggests that reducing long-range interactions using block randomisation, significantly reduces the agreement with the original pair-of-pairs judgements indicating that long-range interactions are important to texture perception.

7.3 Study II

In this study we investigated the effect of removing long-range interactions from images, on pair-of-pairs predictions obtained using the 51 computational feature sets.

We used the 80 "worst" pairs-of-pairs, at each of the six resolutions. We calculated agreement rates between texture features calculated on, and human judgements obtained using, Block Randomised images (AR_R). These were compared against the 80 relevant agreement rates obtained using the original non-randomised texture images (as described in Section 6, i.e., a subset of $POPJ_{POP}$).

Spearman's correlation coefficients [45] computed between the two sets of agreement rates are: -0.035, 0.134, 0.160, 0.224, 0.379 and 0.071 ($P = 0.806, 0.350, 0.275, 0.114, 0.006$ and 0.620) for the resolutions: $1024 \times 1024, 512 \times 512,\ 256 \times 256, 128\times 128, 64 \times 64$, and *Multi*, respectively.

This result suggests that the two sets of agreement rates do not correlate well. Furthermore, compared with the results obtained in Section 6 the Block Randomised agreement rates AR_R, increased on average from $31.3\% \pm 0.1$ to $54.6\% \pm 0.1$. Thus, for this dataset we conclude that humans agree more with computational features when they cannot exploit the long-range interactions originally contained in the textures.

7.4 Summary

The results of Study I suggest that Block Randomisation does affect human perception of long-range interactions, and that it is therefore likely that humans exploit long-range interactions for judging texture similarity. Furthermore, as (1) the majority of the 51 feature sets do not exploit long-range aperiodic interactions and (2) Study II showed that Block Randomisation increased agreement rates between these feature sets and humans, it seems probable that this increase has occurred because we have removed (or at least reduced) long-range interactions. Therefore, we conclude that long-range interactions are important to human estimation of texture similarity.

TABLE 9
Results of Three Kolmogorov-Smirnov (K-S) Tests

K-S Test	Statistic	*df*	Sig. (P)	Is Normal?
AR_B	0.135	10	0.200	Yes
AR_R	0.221	10	0.180	Yes
$AR_B - AR_R$	0.247	10	0.086	Yes

8 Examination of CNN Features

Motivated by the large spatial extents and large numbers of filters that can be exploited by CNNs, we examined the performance of features provided by two *pre-trained* models: VGG-M [24] and VGG-VD-19 [108]. For simplicity, only the pair-of-pairs comparison evaluation method was used (Protocol II).

8.1 CNN Spatial Extent and Statistics

CNNs conveniently divide into two types of processing. Stage I can be considered as the convolutional (Conv) layers. It has been found that some of these layers enable improved predictions of macaque V1 responses to natural images to be provided [21] and they have been found to be selective to 2nd-order statistics [55]. This is not surprising as each layer typically performs convolutional filtering, Relu and optional pooling which can be considered as a specialisation of the popular LNL [81] or "back-pocket" [70] model. Hence, Stage I provides a hierarchy of localised, and at least 2nd-order statistical processing. We use the term "at least" here as positional information is implicitly encoded in these layers.

We consider the fully-connected (FC) layers as Stage II. These can exploit the explicitly localised 2nd-order outputs of the top convolutional layer and its implicit positional coding, to provide processing of potentially image-wide 2nd- and higher order statistical features.

The number of Stage I filters and their maximum spatial extent, for the six different layers of VGG-VD-19 [108] that we used are shown in Table 10. This shows that the effective spatial extent of the Conv features is much greater than those listed in Table 1.

8.2 Conv and FC Features

In order to attribute any improvement in results to the large spatial extent of the Stage I features *or* the potentially image-wide higher order statistics of Stage II, we separately computed Conv and FC features.

Conv features were computed as both global and local means of each feature map extracted from six convolutional layers in VGG-VD-19. The use of global means was motivated by the common structure of conventional texture features [82], [83] and because they reduce the dimensionality of the final feature vectors. Thus the implicit positional information was ignored in "*global mean* Conv features" (i.e., only 2nd-order statistics were exploited). In contrast, for "*local mean* Conv features", means were calculated from *each* spatial cell in a convolutional layer and concatenated to form a single feature vector. Thus the *local mean* Conv feature vectors retain positional information whereas *global mean* Conv feature vectors do not.

The FC features were extracted from the first fully-connected layer: FC6. They therefore had potential to process both image-wide 2nd-order and HOS.

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.

TABLE 10
The Receptive Field Size and Number of Filters of Six Convolutional Layers of VGG-VD-19 [108], and the Agreement Rates (%) Obtained Using the Local Mean and Global Mean Features That are Extracted From These Layers When Compared Against $POPJ_{POP}$

Layer	Conv1_1	Conv2_1	Conv3_1	Conv4_1	Conv5_1	Conv5_4
Receptive Field	3 × 3	10 × 10	24 × 24	68 × 68	156 × 156	**252 × 252**
Num. of Filters	64	128	256	**512**	**512**	**512**
Global Means	51.6	59.2	61.0	69.6	71.8	**73.0**
Local Means	51.5	63.8	62.0	69.5	72.3	**73.2**

Both types of features were L_2 normalised and the *Euclidean* distance were used to compute similarity matrices using the method described in Section 4.1. Since these FC features can only be extracted from 224 × 224 images, 256 × 256 texture images were utilised.

8.3 Performance of the Conv Features

For comparison purposes, we tested six convolutional layers of VGG-VD-19 (Conv1_1, Conv2_1, Conv3_1, Conv4_1, Conv5_1 and Conv5_4) [108]. Agreement rates were calculated for both *global* and *local* mean features using Protocol II and the 1000 $POPJ_{POP}$ judgements [29]. These agreement rates, together with the receptive field size and number of filters for each layer, are reported in Table 10. It shows that the local mean Conv features performed slightly better on average than their global mean counterparts at the cost of using a longer feature vector.

However, the key points arising from these results are (1) that the power of Conv features to estimate texture similarity rises with the size of the receptive field and the number of filters, and (2) that the best single "conventional" feature set (LM [72] which achieved 66.5 percent) and the best Random Forest [18] performance ($67.8 \pm 2.4\%$, mean across three folds) were outperformed by the best CNN features (Conv5_4 at 73.0 and 73.2 percent).

8.4 FC vs. Conv Features

Fully-Connected (FC) agreement rates were computed in the same manner as for Conv features (i.e., Protocol II using $POPJ_{POP}$ [29] ground-truth) but used two pre-trained networks: VGG-M [24] and VGG-VD-19 [108]. They are shown in Table 11 together with agreement rates of: the global mean Conv features; the best single feature set: VAR [91] and best Random Forest [18] result obtained using the same 256 × 256 resolution.

Significance testing CNN and VAR results using McNemar's test [86] shows that the former significantly outperform the latter ($P = 0.00$). Similarly, testing fully-connect (FC) VGG-M results against Random Forest on the same resolution 256 × 256 images (Section 6.7) shows that the former significantly outperforms the latter ($P = 0.00$). There was no significant difference detected between FC and Conv features.

Thus while the best agreement rate (74.3 percent) was achieved using the FC features from network VGG-M, and while this is indicative of promising performances, it is not conclusive that fully-connected versions are better than their purely convolutional counterparts.

However, it is clear that these CNN features outperform their conventional counterparts either when the latter are tested individually or together using a Random Forest [18]. As an illustration of their capability we examined their performance for the four pairs-of-pairs shown in Fig. 10. Each row of this figure shows pairs-of-pairs for which *at least 30* of the conventional feature sets disagreed with the humans, while in contrast, all four CNN feature sets agreed. We attribute the general improved performance of the CNN features to (1) the large receptive fields of: 139 × 139 and 252 × 252 pixels for VGG-M and VGG-VD-19 respectively and (2) the large number of filters employed (512).

TABLE 11
Agreement Rates (%) Obtained Using the CNN [24], [108] Features, VAR [91] and, the Best Random Forest Regressor, Compared With $POPJ_{POP}$ Judgements

FC		Conv (Global Means)		VAR [91]	RF
VGG-M	VGG-VD-19	VGG-M	VGG-VD-19	**256 × 256**	**256 × 256**
74.3	72.8	70.9	73.0	63.2	67.7 ± 2.1

Results for the "FC6" features (derived from the FC6 fully-connected layer) and the global mean "Conv" features (derived from the last convolutional layer) for both of the pre-trained CNNs (VGG-M and VGG-VD-19).

9 Conclusion

We evaluated the ability of 51 popular texture feature sets, computed at multiple resolutions, to estimate two forms of perceptual texture similarity. These ground-truth provide a more stringent assessment of human similarity estimation than is often the case as they were derived using pair-of-pairs [29] and free-grouping [54] experiments. A multi-resolution approach was shown to provide the best overall performance, however, all feature sets were shown to perform significantly less well than the agreement rate obtained between the two types of ground-truth (73.9 percent). When the 51 feature sets were combined using a Random Forest [18] approach, results improved to a best average of 67.8 percent, however this was still significantly below the agreement rate of the two human-derived ground-truth sets.

As analysis of these feature sets showed that few exploit long-range HOS, we investigated the effect of removing long-range interactions from textures. The results showed that the computational features agreed significantly better with human observers when long-range interactions were removed. We therefore conclude:

1) that long-range interactions are important for human judgement of texture similarity (this is reinforced by Sharan *et al.* [107] who showed that it was difficult for humans to identify material categories when shown globally scrambled but locally preserved images) and
2) that the computational features do not exploit these data well.

In contrast to most conventional features, Convolutional Neural Networks (CNNs) provide the capability to learn large numbers of filters from large datasets and compute image-wide statistics of almost arbitrary order. Given the size of our two ground-truth sets we elected to investigate the performances of two *pre-trained* CNNs [24], [108]. Such networks provide two types of features: *Conv* features derived from the convolutional layers (which can provide estimates of 2nd-order statistics over relatively large spatial extent) and *FC* features derived from the first fully-connected layer

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.

(which provide potentially image-wide 2nd- *and* higher order statistics). Our results showed that both the Conv and the FC features provided superior performance to the conventional features whether these conventional features were used individually or in combination in the Random Forest [18]. This is a complementary result to that of Zhang *et al.* [133] which showed that CNN-based approaches improved estimation of human perception of image patch distortion.

In summary our findings are as follows.

1) Using fine-grained perceptual texture similarity ground-truth we showed that a multi-resolution approach provides significantly higher retrieval scores, but that no overall "best" feature set could be discerned.
2) Using 51 feature sets in a Random Forest we showed that this approach significantly improved results over using individual feature sets, suggesting that variety of feature is important.
3) As others have found [46], [73], [98], [113] we established, using Block Randomisation, that long-range interactions are important to human perception of texture similarity, but that many of the conventional computational features tested do not exploit long-range HOS.
4) We showed that using CNN features derived from two pre-trained networks provided significantly better results than either (1) or (2). This is most likely due to the large number of pre-trained filters employed and the potentially image-wide spatial support of features derived from the top layers.

Acknowledgments

X. Dong completed the majority of this research when he worked with the Texture Lab, Heriot-Watt University, Edinburgh, UK. The authors would like to thank the editor and anonymous reviewers for their extensive comments and help in improving this paper. They acknowledge the support of Heriot-Watt University.

References

[1] 1995. [Online]. Available: http://vismod.media.mit.edu/vismod/imagery/VisionTexture/
[2] 2003. [Online]. Available: http://www.macs.hw.ac.uk/texturelab/resources/databases/photex/
[3] 2011. [Online]. Available: http://www.macs.hw.ac.uk/texturelab/resources/databases/pertex/
[4] 2004. [Online]. Available: http://www.nada.kth.se/cvap/databases/kth-tips/
[5] 2009. [Online]. Available: http://www.wavelab.at/sources/STex/
[6] 2005. [Online]. Available: http://www-cvr.ai.uiuc.edu/ponce_grp/data/
[7] Accessed: 2018. [Online]. Available: www.colour-affects.co.uk/psychological-properties-of-colours
[8] N. Abbadeni, "Computational perceptual features for texture representation and retrieval," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 236–246, Jan. 2011.
[9] F. Ade, "Characterisation of texture by 'eigenfilter'," *Signal Process.*, vol. 5, pp. 451–457, 1983.
[10] T. Ahonen, J. Matas, C. He, and M. Pietikainen, "Rotation invariant image description with local binary pattern histogram fourier features" in *Proc. Scand. Conf. Image Anal.*, 2009, pp. 61–70.
[11] T. Ahonen and M. Pietikäinen, "Image description using joint distribution of filter bank responses," *Pattern Recognit. Lett.*, vol. 30, no. 4, pp. 368–376, 2009.
[12] M. Amadasun and R. King, "Textural features corresponding to textural properties," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, no. 5, pp. 1264–1274, Sep./Oct. 1989.
[13] J. Bar-Ilan, K. Keenoy, E. Yaari, and M. Levene, "User rankings of search engine results," *J. Am. Society Inf. Sci. Technol.*, vol. 58, no. 9, pp. 1254–1266, 2007.
[14] G. Berger and R. Memisevic, "Incorporating long-range consistency in CNN-based texture generation," *arXiv:1606.01286v2*.
[15] U. Bergmann, N. Jetchev, and R. Vollgraf, "Learning texture manifolds with the periodic spatial GAN" in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 469–477.
[16] A. Borji, D.N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 55–69, Jan. 2012.
[17] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localised spatial filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 55–73, Jan. 1990.
[18] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[19] P. Brodatz, *Textures: A Photographic Album for Artists and Designers.* Mineola, NY, USA: Dover Publications, 1966.
[20] P. J. Burt, "Fast filter transforms for image processing," *Comput. Graphics Image Process.*, vol. 16, no. 1, pp. 20–51, 1981.
[21] S. A. Cadena, *et al.*, "Deep convolutional models improve predictions of macaque v1 responses to natural images," *PLoS Comput. Biol.*, vol. 15, no. 4, 2019. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1006897
[22] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
[23] M. Chantler, M. Petrou, A. Penirsche, M. Schmidt, and G. MGunnigle, "Classifying surface texture while simultaneously estimating illumination direction" *Int. J. Comput. Vis.*, vol. 62, no. 1-2, pp. 83–96, 2005.
[24] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional networks," in *Proc. British Mach. Vis. Conf.*, M. Valstar, A. French, and T. Pridmore, Eds., BMVA, 2014. [Online]. Available: http://dx.doi.org/10.5244/C.28.6
[25] B. B. Chaudhuri, N. Sarkar, and P. Kundu, "Improved fractal geometry based texture segmentation technique," *IEE Proc. Comput. Digital Techn.*, vol. 140, no. 5, pp. 233–241, Sep. 1993.
[26] R. Chellappa and S. Chatterjee, "Classification of textures using gaussian markov random fields," *IEEE Trans. Acoustics Speech Signal Process.*, vol. ASSP-33, no. 4, pp. 959–963, Aug. 1985.
[27] R. Y. Cho, V. Yang, and P. E. Hallett, "Reliability and dimensionality of judgements of visually textured materials," *Perception Psychophysics*, vol. 62, no. 4, pp. 735–752, 2000.
[28] M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi, "Deep filter banks for texture recognition, description, and segmentation," *Int J. Comput. Vis.*, vol. 118, no. 1, pp. 65–94, 2016.
[29] A. D. F. Clarke, X. Dong, and M. J. Chantler, "Does free-sorting provide a good estimate of visual similarity," in *Proc. 3rd Int. Conf. Appearance Predicting Perceptions*, 2012, pp. 17–20.
[30] A. D. F. Clarke, P. R. Green, F. Halley, and M. J. Chantler, "Similar symmetries: The role of wallpaper groups in perceptual texture similarity," *Symmetry*, vol. 3, no. 2, pp. 246–264, 2011.
[31] J. M. Coggins and A. K. Jain, "A spatial filtering approach to texture analysis," *Pattern Recognit. Lett.*, vol. 3, pp. 195–203, 1985.
[32] M. Crosier and L. D. Griffin, "Using basic image features for texture classification," *Int. J. Comput. Vis.*, vol. 88, pp. 447–460, 2010.
[33] K. J. Dana, G. Medioni, and S. Dickinson, *Comput. Texture Patterns: From Textons Deep Learn.*, San Rafael, CA, USA: Morgan & Claypool, 2018.
[34] K. J. Dana, B. Van Ginneken, S. K. Nayar, and J. J. Koenderink, "Reflectance and texture of real world surfaces," *ACM Trans. Graphics*, vol. 18, no. 1, pp. 1–34, 1999.
[35] H. A. David, *The Method of Paired Comparisons*. New York, NY, USA: Oxford Univ. Press, 1988.
[36] X. Dong, "Perceptual texture similarity estimation," PhD thesis, Dept. School Math. Comput. Sci., Heriot-Watt University, Edinburgh, United Kingdom, 2014.
[37] X. Dong and M. J. Chantler, "The importance of long-range interactions to texture similarity," in *Proc. 15th Int. Conf. Comput. Anal. Images Patterns*, 2013, vol. 8047, pp. 425–432.

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.

[38] X. Dong, T. Methven, and M. J. Chantler, "How well do computational features perceptually rank textures? A comparative evaluation," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2014, pp. 281–288.

[39] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2016, pp. 658–666.

[40] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. 28th Annu. Conf. Comput. Graphics Interactive Techn.*, 2001, pp. 341–346.

[41] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: benchmark and bag-of-features descriptors," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 11, pp. 1624–1636, Nov. 2011.

[42] K. Emrith, M. J. Chantler, P. R. Green, L. T. Maloney, and A. D. F. Clarke, "Measuring perceived differences in surface texture due to changes in higher order statistics," *J. Optical Society America A*, vol. 27, no. 5, pp. 1232–1244, 2010.

[43] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top K lists," in *Proc. ACM-SIAM Symp. Discrete Algorithms-SODA*, 2003, vol. 17, pp. 28–36.

[44] S. Fan, T. Ng, J.S. Herberg, B.L. Koenig, C. Y. -C. Tan, and R. Wang, "An automated estimator of image visual realism based on human cognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4201–4208.

[45] A. Field, *Discovering Statistics Using SPSS*, Thousand Oaks, CA, USA: SAGE Publications, 2009.

[46] D. J. Field, A. Hayes, and R. F. Hess, "Contour integration by the human visual system: evidence for a local 'association field'," *Vis. Res.*, vol. 33, pp. 173–193, 1993.

[47] J. Filip and M. Haindl, "Bidirectional texture function modeling: A state of the art survey," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 31, no. 11, pp. 1921–1940, Nov. 2009.

[48] I. Fogel and D. Sagi, "Gabor filters as texture discriminator," *Biol. Cybern.*, vol. 61, pp. 103–113, 1989.

[49] C. Fraley and A. E. Raferty, "How many clusters? which clustering method? answers via model-based cluster analysis?," *Comput. J.*, vol. 41, no. 8, pp. 578–588, 1998.

[50] K. Fujii, S. Sugi, and Y. Ando, "Textural properties corresponding to visual perception based on the correlation mechanism in the visual system," *Psychol. Res.*, vol. 67, no. 3, pp. 197–208, 2003.

[51] M. M. Galloway, "Texture classification using gray level run lengths," *Comput. Graphics Image Process.*, vol. 4, pp. 172–179, 1975.

[52] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 262–270.

[53] S.W. Greenhouse and S. Geisser, "On methods in the analysis of profile data," *Psychometrika*, vol. 24, pp. 95–112, 1959.

[54] F. Halley, "Perceptually relevant browsing environments for large texture databases," PhD thesis, Heriot Watt Univ., Edinburgh, 2011.

[55] L. E. Hallum, M. S. Landy, and D. J. Heeger, "Human primary visual cortex (V1) is selective for second-order spatial frequency," *J. Neurophysiol*, vol. 105, pp. 2121–2131, 2011.

[56] R. M. Haralick, "Statistical and structural approaches to texture," in *Proc. IEEE*, vol. 67, no. 5, pp. 786–804, May 1979.

[57] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[58] N. Hariri, "Relevance ranking on Google: Are top ranked results really considered more relevant by the users?," *Online Inf. Rev.*, vol. 35, no. 4, pp. 598–610, 2011.

[59] D. Harwooda, T. Ojalab, M. Pietikäinen, S. Kelmanc, and L. Davisa, "Texture classification by center-symmetric auto-correlation, using Kullback discrimination of distributions," *Pattern Recognit. Lett.*, vol. 16, no. 1, pp. 1–10, 1995.

[60] C. Heaps and S. Handel, "Similarity and features of natural textures," *J. Exp. Psychol.: Human Perception Perform.*, vol. 25, pp. 299–320, 1999.

[61] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using gabor filters," *Pattern Recognit.*, vol. 24, pp. 1167–1186, 1991.

[62] F. B. Joanl, "Guinness, gosset, fisher, and small samples," *Statist. Sci.*, vol. 2, no. 1, pp. 45–52, 1987.

[63] B. Julesz, "Textons, the elements of texture perception, and their interactions," *Nature*, vol. 290, no. 5802, pp. 91–97, 1981.

[64] A. Kadyrov and M. Petrou, "The trace transform and its applications," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 23, no. 8, pp. 811–828, Aug. 2001.

[65] F. Khelifi and J. Jiang, "k-NN regression to improve statistical feature extraction for texture retrieval," *IEEE Trans. Image Precess.*, vol. 20, no. 1, pp. 293–298, Jan. 2011.

[66] J. K. Kim and H. W. Park, "Statistical textural features for detection of microcalcifications in digitised mammograms," *IEEE Trans. Medical Imaging*, vol. 18, no. 3, pp. 231–238, Mar. 1999.

[67] J. J. Koenderink, "The structure of images," *Biol. Cybern.*, vol. 50, no. 5, pp. 363–370, 1984.

[68] P. J, Kohler, A. Clarke, A. Yakovleva, Y. Liu, and A. M. Norcia, "Representation of maximally regular textures in human visual cortex," *J. Neurosci.*, vol. 36, no. 3, pp. 714–729, 2016.

[69] A. N. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," *Giornale dell'Istituto Italiano degli Attuari*, vol. 4, pp. 83–91, 1933.

[70] M. S. Landy and N. Graham, "Visual perception of texture," *Vis. Neurosciences*, vol. 2, pp. 1106–1118, 2004.

[71] K. I. Laws, "Rapid texture identification," in *Proc. SPIE Conf. Image Process. Missile Guidance*, 1980, pp. 376–380.

[72] T. Leung and J. Malik, "Representing and recognising the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vis.*, vol. 43, pp. 29–44, 2001.

[73] W. C. Lin, J. Hays, C. Wu, V. Kwatra, and Y. Liu, "Quantitative evaluation on near regular texture synthesis algorithms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 427–434.

[74] W. Lin and Y. Liu, "A lattice-based MRF model for dynamic near-regular texture tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 5, pp. 777–792, May 2007.

[75] L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, and M. Pietikäinen, "From BoW to CNN: Two decades of texture representation for texture classification," *Int. J. Comput. Vis.*, vol. 127, no. 1, pp. 74–109, 2019.

[76] Y. Liu, W. Lin, and J. Hays, "Near-regular texture analysis and manipulation," *ACM Trans. Graphics*, vol. 23, no. 3, pp. 368–376, 2004.

[77] Z.Q. Liu and S.V.R. Madiraju, "Covariance-based approach to texture processing," *Appl. Optics*, vol. 35, no. 5, pp. 848–853, 1996.

[78] Y. Liu, Y. Tsin and W. Lin, "The promise and perils of near-regular texture," *Int. J. Comput. Vis.*, vol. 62. No. 1-2, pp. 145–159, 2005.

[79] H. Long, W. K. Leow, and F. K. Chua, "Perceptual texture space for content-based image retrieval," in *Proc. Int. Conf. Multimedia Model.*, 2000, pp. 167–180.

[80] D. G. Lowe, *Perceptual Organisation and Visual Recognition*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1985.

[81] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanisms," *J. Optical Soc. America*, vol. 7, no. 5, pp. 923–932, 1990.

[82] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 18, no. 8, pp. 837–842, Aug. 1996.

[83] J. Mao and A. K. Jain, "Texture classification and segmentation using multiresolution simultaneous autoregressive models," *Pattern Recognit.*, vol. 25, no. 2, pp. 173–188, 1992.

[84] T. Matthews, M. S. Nixon, and M. Niranjan, "Enriching texture analysis with semantic data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1248–1255.

[85] J. W. Mauchly, "Significance test for sphericity of a normal n-variate distribution," *Ann. Math. Statist.*, vol. 11, no. 2, pp. 204–209, 1940.

[86] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[87] P. T. Metaxas, L. Ivanova, and E. Mustafaraj, "New quality metrics for web search results," in *Proc. 4th Int. Conf. Web Inf. Syst. Technol.*, 2009, pp. 278–292.

[88] K. Mikolajczyk *et al.*, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, pp. 43–72, 2005.

[89] I. Ng, T. Tan, and J. Kittler, "On local linear transform and gabor filter representation of texture," in *Proc. Int. Conf. Pattern Recognit.*, 1992, pp. 627–631.

[90] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognit.*, vol. 29, pp. 51–59, 1996.

[91] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution greyscale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.

[92] V. Ojansivu, E. Rahtu, and J. Heikkila, "Rotation invariant local phase quantisation for blur insensitive texture analysis," in *Proc. Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.

[93] B. A. Olshausen and D. J. Field, "Natural image statistics and efficient coding," *Netw.: Comput. Neural Syst.*, vol. 7, no. 2, pp. 333–339, 1996.

[94] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proc. IEEE*, vol. 69, no. 5, pp. 529–541, May 1981.

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

[96] M. Parseval, "Mémoire sur les séries et sur l'intégration complète d'une équation aux différences partielles linéaire du second ordre, à coefficients constants," *Mémoires présentés à l'Institut des Sciences, Lettres et Arts, par divers savans, et lus dans ses assemblées. Sciences, mathématiques et physiques. (Savans étrangers)*, vol. 1, pp. 638–648, 1806.

[97] J. S. Payne, L. Hepplewhite, and T. J. Stonham, "Perceptually based metrics for the evaluation of textural image retrieval methods," in *Proc. IEEE Int. Conf. Multimedia Comput. Syst.*, 1999, vol. 2, pp. 793–797.

[98] U. Polat, "Functional architecture of long-range perceptual interactions," *Spatial Vis.*, vol. 12, pp. 143–162, 1999.

[99] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int. J. Comput. Vis.*, vol. 40, no. 1, pp. 49–71, 2000.

[100] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C*, 2nd ed., Cambridge, U.K.: Cambridge Univ. Press, 1992.

[101] E. Rahtu, M. Salo, and J. Heikkila, "Affine invariant pattern recognition using multiscale autoconvolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 908–918, Jun. 2005.

[102] T. Randen and J. H. Husøy, "Filtering for texture classification: A comparative study," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 21, no. 4, pp. 291–310, Apr. 1999.

[103] A. R. Rao and G. L. Lohse, "Identifying high level features of texture perception," *Graphical Models Image Process.*, vol. 55, no. 3, pp. 218–233, 1993.

[104] T. Reed and J. Buf, "A review of recent texture segmentation and feature extraction techniques," *Comput. Vis. Image Process. Graphics*, vol. 57, no. 3, pp. 359–372, 1993.

[105] S. Santini and R. Jain, "Similarity measures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 9, pp. 871–883, Sep. 1999.

[106] C. Schmid, "Constructing models for content-based image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogni.*, 2001, vol. 2, pp. 39–45.

[107] L. Sharan, C. Liu, R. Rosenholtz, and E. H. Adelson, "Recognizing materials using perceptually inspired features," *Int. J. Comput. Vis.*, vol. 103, pp. 348–371, 2013.

[108] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale visual recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.

[109] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1470–1477.

[110] N. Smirnov, "Tables for estimating the goodness of fit of empirical distributions," *Ann. Math. Statist.*, vol. 19, pp. 279–281, 1948.

[111] G. Smith and I. Burns, "Measuring texture classification algorithms," *Pattern Recognit. Lett.*, vol. 18, no. 14, pp. 1495–1501, 1997.

[112] I. Sobel, "An isotropic 3 × 3 gradient operator," *Machine Vision for Three-Dimensional Scenes*, H. Freeman eds., Cambridge, MA, USA: Academic Press, 1990, pp. 376–379.

[113] L. Spillmann and J. S. Werner, "Long-range interactions in visual perception," *Trends Neurosciences*, vol. 19, pp. 428–434, 1996.

[114] H. Tamura, S. Mori, and T. Yamawaki, "Textural features corresponding to visual perception," *IEEE Trans. Syst., Man, Cybern.*, vol. 8, no. 6, pp. 460–473, Jun. 1978.

[115] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 500, pp. 2319–2323, 2000.

[116] A. Toet, "Computational versus psychophysical bottom-up image saliency: A comparative evaluation study," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 33, no. 11, pp. 2131–2146, Nov. 2011.

[117] M. Tuceryan and A. K. Jain, "Texture analysis," *Handbook of Pattern Recognition and Computer Vision*, C. H. Chen, L. F. Pau, and P.S.P. Wang, eds., Singapore: World Scientific, 1993, pp. 235–276.

[118] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky, "Texture networks: feed-forward synthesis of textures and stylized images," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 1349–1357.

[119] M. Unser, "Sum and difference histograms for texture classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 1, pp. 118–125, Jan. 1986.

[120] I. Ustyuzhaninov, W. Brendel, L. Gatys, and M. Bethge, "What does it take to generate natural textures?" in *Proc. Int. Conf. Learn. Representations*, 2017.

[121] L. Van Gool, P. Dewaele, and A. Oosterlinck, "Texture analysis Anno 1983," *Comput. Vis. Graphics Image Process.*, vol. 29, no. 3, pp. 336–357, 1985.

[122] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int. J. Comput. Vis.*, vol. 62, pp. 61–81, 2005.

[123] M. Varma and A. Zisserman, "A statistical approach to material classification using image patch exemplars," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 31, no. 11, pp. 2032–2047, Nov. 2009.

[124] F. M. Vilnrotter, R. Nevatia, and K. E. Price, "Structural analysis of natural textures," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. TPAMI-8, no. 1, pp. 76–89, Jan. 1986.

[125] X. Wang, F. Albregtsen, and B. Foyn, "Texture features from gray level gap length matrix," *Proc. IAPR Int. Conf. Mach. Vis. Appl.*, 1994, pp. 375–378.

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

[127] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. Signals Syst. Comput.*, 2003, pp. 1398–1402.

[128] R. Wenger, "Visual art, archaeology and gestalt," *Leonardo*, vol. 30, pp. 35–46, 1997.

[129] J. S. Weszka, C. R. Dyer, and A. Rosenfeld, "A comparative study of texture measures for terrain classification," *IEEE Trans. Syst., Man, Cybernet.*, vol. SMC-6, no. 4, pp. 269–285, Apr. 1976.

[130] X. Xie and M. Mirmehdi, "A galaxy of texture features," *Handbook of Texture Analysis*, M. Mirmehdi, X. Xie, and J. Suri, eds., Singapore: World Scientific, 2009.

[131] L. Ying, "Phase unwrapping," *Wiley Encyclopedia Biomed. Eng.*, vol. 6, pp. 1–11, 2006.

[132] B. Zhang, Y. Gao, S. Zhao, and J. Liu, "Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 533–544, Feb. 2010.

[133] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogniti.*, 2018, pp. 586–595.

[134] H. Zhang, J. Xue, and K. Dana, "Deep TEN: Texture encoding network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2896–2905.

[135] S. Zhu, "Statistical modeling and conceptualization of visual patterns" *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 25, no. 6, pp. 691–712, Jun. 2003.

[136] J. Zujovic, "Perceptual texture similarity metrics" PhD thesis, Dept. Electr. Eng. Comput. Sci., Northwestern Univ., Evanston, IL, 2011.

Xinghui Dong received the PhD degree from Heriot-Watt University, United Kingdom, in 2014. He is currently working as a research associate with the Centre for Imaging Sciences, The University of Manchester, United Kingdom. His research interests include automatic defect detection, image representation, texture analysis, and visual perception.

Junyu Dong received the BSc and MSc degrees from the Ocean University of China, in 1993 and 1999, respectively, and the PhD degree in image processing from Heriot-Watt University, in 2003. He joined the Ocean University of China, in 2004, and is currently a professor and the head of the Department of Computer Science and Technology. His research interests include machine learning, big data, computer vision and underwater image processing.

Mike J. Chanter received the PhD degree, in 1994, studying the effect of illumination direction on image texture. Since then, he has studied many aspects of three-dimensional surface texture and founded the Texture Lab at Heriot-Watt. He has supervised more than 20 PhDs, published more than 120 papers, and attracted funding from EU and United Kingdom research directorates in it and the related areas.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.

Authorized licensed use limited to: University of Ioannina. Downloaded on May 16,2022 at 12:11:00 UTC from IEEE Xplore. Restrictions apply.