Perceptually Inspired Normalized Conditional Compression Distance

Nima Nikvand & Zhou Wang
Department of Electrical & Computer Engineering
University of Waterloo
Waterloo, Ontario, Canada
nnikvand@uwaterloo.ca
zhou.wang@uwaterloo.ca

Wisam Farjow & Xavier Fernando & S. Younes Sadat-Nejad
Department of Electrical & Computer Engineering
Ryerson University
Toronto, Ontario, Canada
wfarjow@hotmail.com
fernando@ee.ryerson.ca
seyedyouns.sadatneja@ryerson.ca

***Abstract*—Image similarity measurement is a common issue in a broad range of applications in image processing, recognition, classification and retrieval. Conventional image similarity measures are often limited to specific applications and cannot be applied in general scenarios. The theory of Kolmogorov complexity provides a universal framework for a generic similarity metric based on information distance between objects. Normalized Information Distance (NID) has been shown to be a valid and universal distance metric applicable in measurement of similarity of any two objects, and has been successfully applied to a wide range of applications in the past. The difficulty of NID lies in the non-computable nature of the Kolmogorov complexity, and thus approximation has to be applied in practice. Here we propose a perceptually-inspired Normalized Conditional Compression Distance (NCCD) measure by using the Divisive Normalization Transform (DNT) as a means to model the non-linear behavior of the Human Visual System (HVS) in reducing statistical dependencies of visual signals for efficient representation, and show that this perceptual extension of NID can be used in a wide range of image processing applications, including texture classification and face recognition.**

***Index Terms*—mage Similarity Measurement; Kolmogorov complexity; Normalized Information Distance; Compression; Divisive Normalization; Classification; Recognition mage Similarity Measurement; Kolmogorov complexity; Normalized Information Distance; Compression; Divisive Normalization; Classification; Recognition I**

I. Introduction

Measuring the similarity between images is a fundamental problem to many applications throughout the entire field of image processing and machine vision. These applications include analysis of texture, image classification, pattern recognition and detection and tracking of objects. Conventional image similarity measures are limited to specific applications, and the existing "general-purpose" methods cannot be applied in many scenarios. Image similarity measures have seen significant progress in recent years [1]. For example the Structural Similarity Index (SSIM) has been shown to perform much better in predicting visual quality of images compared to traditional and widely used measures such as Mean Squared Error (MSE) [2], [3]. Despite its success in similarity and quality assessment of images, SSIM is very sensitive to small geometric distortions such as shifting, scaling, and rotation [4]–[6]. By contrast, the human visual system (HVS) is very resilient towards this type of distortions, and attempts to describe geometric changes among similar images by finding the shortest description for the change. This might be exemplified by an image of 8×8 checkerboard and another image of the same checkerboard but with the black and white squares exchanged. The Human mind quickly notices this change and finds the two image similar, however many existing image similarity measurement methods fail to recognize this similarity. When asked to describe the difference between the two images, a simple description is usually given, i.e. to flip the color of all squares. The Kolmogorov complexity measure targets at a more fundamental solution to the problem. Given the first image, Kolmogorov complexity looks for the shortest program to produce the second one on a Universal Turing Machine (UTM), and flipping the bits in the first checkerboard image is likely a good approximate of the Kolmogorov complexity.

The theory of Kolmogorov complexity provides solid groundwork to build a universal and generic distance metric between objects which minorizes all metrics in its class. Normalized Information Distance (NID) has been shown to be a valid and universal distance metric applicable in measurement of similarity of any two objects, and has been successfully applied to a wide range of applications in the past [7]. The difficulty in using this distance in practice lies in the non-computable nature of Kolmogorov complexity.

Due to the fact that the notion of Kolmogorov complexity treats all the bits in a program equally, all NID based frameworks do not take into account the degree of perceptual relevance of the information contained in the image. The HVS on the other hand is adapted to match statistical properties of natural stimuli [8]. It is hypothesized that the early sensory systems remove redundancy in the stimuli which results in a set of statistically independent neural responses [9]. In this sense, it is necessary for HVS to filter visual stimuli, during which the information in the stimuli signal is not treated equally. In order to account for perceptual relevance of the bits in the approximation of Kolmogorov complexity, a theoretical decomposition of NID to perceptually relevant information and residue was proposed [10]. The model is based on deriving a vector of relevant information which can be used to reconstruct a lossy representation of the image data. The method however does not propose a practical framework for this decomposition, and leaves the feature extraction and selection process to ad-hoc, application specific algorithms.

A practically more useful technique in removing redundancies in the stimuli signal is by using efficient coding transforms. The advantage of using such models is to reduce the perceptual and statistical dependence of stimuli and to represent perceptually relevant information of the signal most efficiently [9]. Among many nonlinear efficient coding transforms, Divisive Normalization Transform (DNT) [9] has been extensively studied in the past. It has been observed that this simple normalization of elements by a weighted Minkowski combination of its neighboring elements can significantly reduce the dependencies among elements of natural image [11].

In this study, we extend the NID framework by introducing the concept of Divisive Normalization Transform (DNT) as a model for the nonlinear behavior of the HVS in removing redundancies from the natural scene images based on the efficient coding principle [8], [9]. The robustness and simpicility of the proposed perceptually inspired Normalized Conditional Compression Distance (NCCD) framework allows us to use this method in diverse applications such as image classification, retrieval and recognition without the need to train the algorithm on large datasets or set any domain-specific parameters. We demonstrate the perceptual NCCD framework to texture classification and face recognition problems, and compare the results to existing compression based [12] and sparsity based [13] similarity methods in the literature.

The goal in this study is to introduce a generic image similarity measurement method which can be applied to various scenarios such as image classification and clustering, face recognition and retrieval. The experimental results reported here focus on demonstrating the wide applicability of the proposed method, which can be readily applied to new application without adapting any domain specific knowledge and without any training or parameter tuning process. Therefore, the method is compared to similar information distance based algorithms with generic applicability.

II. Normalized Conditional Compression Distance (NCCD) Framework

The Kolmogorov complexity [14] of an object is defined to be the length of the shortest program that can produce that object on a universal Turing machine and halt:

$$K(x) = \min_{p:U(p)=x} l(p), \tag{1}$$

where K is the Kolmogorov complexity of the object x, and l is the length of the program p which produces object x on the Universal Turing Machine (UTM) U. In [7], the authors assume the existence of a general decompressor that can be used to decompress the presumably shortest program x^* to the desired object x.

The conditional Kolmogorov complexity of x relative to y is denoted by $K(x|y)$. An information distance between x and y can then be defined as $\max\{K(x|y), K(y|x)\}$, which is the maximum of the length of the shortest program that computes x from y and y from x. To convert it to a normalized symmetric metric, an NID measure was introduced in [7]:

$$\mathrm{NID}(x,y) = \frac{\max\{K(x|y^*), K(y|x^*)\}}{\max\{K(x), K(y)\}}. \tag{2}$$

It was proved that NID is a valid *distance metric* that satisfies the identity and symmetry axioms and the triangular inequality [7].

The real-world application of NID is difficult because Kolmogorov complexity is a non-computable quantity [14]. By using the fact that $K(xy) = K(y|x^*) + K(x) = K(x|y^*) + K(y)$ (subject to a logarithmic term), and by approximating Kolmogorov complexity K using a practical data compressor, a Normalized Compression Distance (NCD) was proposed in [7] as

$$\mathrm{NCD}(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}, \tag{3}$$

where C is any general-purpose data compressor, e.g. CALIC [15]. NCD has been proved to be an effective approximation of NID and achieves superior performance in certain applications such as the construction of phylogeny trees using DNA sequences [7]. Nevertheless, when NCD was used to quantify image similarities, it did not achieve the same level of success. For example, it was reported in [16] that NCD works well when parts are added or subtracted from an image, but struggles when image variations involve form, material and structure. We believe that one main reason is the poor approximation of $K(xy)$ using $C(xy)$, which is often implemented by applying a regular image compressor to the concatenation of two images. For example, when an image is a ninety-degree rotated copy of another, concatenating two images would not facilitate any efficient compression. To avoid this problem, we propose to approximate the conditional Kolmogorov complexity in Eq. (2) directly by designing a conditional image compressor denoted by C_T, so that

$$K(y|x) \approx C_T(y|x) \quad \text{and} \quad K(x|y) \approx C_T(x|y)\,. \tag{4}$$

This leads to an NCCD measure [17] given by

$$\mathrm{NCCD}(x,y) = \frac{\max\{C_T(x|y), C_T(y|x)\}}{\max\{C(x), C(y)\}}. \tag{5}$$

It remains to define the conditional compressor C_T. Here we propose a practical solution by making use of a set of transformations that convert one image to another. Let $\{T_i|i = 1, \cdots, N\}$ be the set of transformations, let $T_i(x)$ represent the transformed image when applying the i-th transform to image x, and let $p(T_i, x)$ denote the parameters used in the transformation. Each type of transformation is also associated with a parameter compressor, and C_i^p denotes the parameter compressor of the i-th transformation. We can then define our conditional compressor as

$$C_T(y|x) = \min_i\{C[y - T_i(x)] + C_i^p[p(T_i, x)] + \log_2(N)\}, \tag{6}$$

where C remains to be a practical image compressor which encodes the difference between y and the transformed image $T_i(x)$, and the $\log_2(N)$ term computes the number of bits required to encode the selection of one out of N potential transformations.

The idea of finding the simplest transformation between two images is sensible from the viewpoint of human visual perception, for which it has long been hypothesis that the biological visual system is an efficient coder of the visual world [8]. For example, given two images that are rotated copy of each other, our visual system would not interpret the difference between them by directly differencing their intensity values (which requires a large number of bits to encode the residual), but by estimating the amount of rotation (which can be coded very efficiently).

An advantage of NCCD (as opposed to NCD) is that it provides a more flexible framework so that different types of transformations can be included. The list of transformations can also be incremental, in the sense that new transformations, when available, can be easily added into the existing list, and expanding the list always improves the approximation of NCCD to NID. Of course, exhausting all possible transformations is practically impossible. However, by going through a handful of transformations, it may be sufficient to appropriately cover most image distortions encountered in real-world applications.

III. Implementation of Divisive Normalization Transform Based NCCD

A. Divisive Normalization Transform and Perceptuality

Biological sensory systems are believed to have evolved to match the statistical properties of natural stimuli, and efficient coding principle provides a powerful explanation for such an evolutionary optimization by asserting that sensory systems represent information content of the stimuli subject to their inherent limitations [9], [18]. Visual stimulus in this context is any natural image that stimulates the optic nerve. The stimuli information passes through the optics of the eyes and is transmitted to the brain via the optic nerve using electrical impulses, often referred to as action potentials or signals. It is hypothesized that the signals in the visual system form a neural code for efficiently representing the sensory information [8]. In order to model this mechanism in the mammalian visual cortex, a set of sensory transforms have been proposed in the past, all of which attempt to reduce statistical dependencies of stimuli signals for more efficient signal transmission, representation, and processing [9]. The most simple form of these transforms which was originally proposed to model non-linearities in neurons of visual cortex [19] is Divisive Normalization (DN). Divisive Normalization is often modeled as dividing the value of a pixel divided by a weighted average of values of adjacent pixels in spatial domain which is directly corresponding to a neural response model of HVS. Previous studies have shown that DN can reduce statistical dependencies among sensory signals [9], [20], act as a maximum likelihood estimator in noiseless data estimation [21], and play an important role in general decision making mechanism in context-dependent scenarios [22].

B. Novel Perceptual NCCD

Inspired by the redundancy reduction properties of divisive normalization, we use an energy based form of DNT to create a perceptually uniform conditional image in the spatial domain. Assuming that the target image X is given in the approximation of an upper bound for conditional Kolmogorov complexity of the source image Y, we propose to normalize the conditional image by the following local pooling map:

$$T_0\sqrt{1+\frac{\sigma_x^2}{C_0}} \tag{7}$$

where T_0 and C_0 are constants and σ_x^2 is the local energy of the given image computed using an 11×11 sliding Gaussian window with variance of 1.5. This normalization follows the same logic as DNT in the spatial domain, and helps get rid of the perceptual redundancies in the conditional image and results in significant improvement in the compressed file size and therefore complexity of the conditional image. Since the image X is presumed to be available to the decoding machine, the size of this divisive normalization map is not required to be encoded in the approximation of conditional Kolmogorov complexity. Furthermore, the proposed transform can be considered as a lossy compression scheme, and with the proper choice for C_0 and T_0, the resulting normalized image can be used to reconstruct the original image with high structural similarity with the original image.

Our current implementation of NCCD are as follows. First, we adopt the content adaptive lossless image compression algorithm (CALIC) [15] as the base image compressor, which achieves superior performance when compared with state-of-the-art algorithms. CALIC is employed in computing the denominator of Eq. (5) as well as the first term in Eq. (6). Since $y - T_i(x)$ in Eq. (6) can generate negative values and CALIC applies to grayscale images with positive intensity values only, the mean intensity value of $y - T_i(x)$ is shifted to mid-gray level before the application of CALIC. Second, the types of transformations involved in the computation of C_T include:

1) *Global contrast and luminance change*. This is computed by a point-wise intensity transformation defined as $s = \alpha(r - \bar{r}) + \bar{r} + \beta$, where r and s are the intensity values before and after the transformation, respectively, $\bar{r}$ is the average value of r, and α and β are the parameters that determine the degrees of contrast and mean luminance changes, respectively. In a special case when α = 1 and β = 0, it reduces to an identity transform, i.e., $T(x) = x$.
2) *Global Fourier power spectrum scaling*. This transformation attempts to match two images by scaling the power spectrum of one image in the Fourier transform domain. Let $X(\omega)$ and $Y(\omega)$ be the Fourier transforms of x and y, respectively. We first find the best linear transform parameters p_1 and p_2, such that $\||Y(\omega)| -$

$(p_1|X(\omega)| + p_2)\|^2$ is minimized. We then define the transform $T(x)$ as the inverse Fourier transform of $p_1 X(\omega) + p_2$.

3) *Global affine transform.* This transformation tries to match one image by applying a global affine transform to another. The transformation can be encoded using six parameters and covers a variety of image changes including translation, scaling (zooming in or zooming out), rotation, and shearing.
4) *Local registration transformation.* This is implemented by aligning two images using the local affine and global smooth registration [23].

Given a pair of images x and y for comparison, we attempt all the above transformations from both x to y and y to x (multiple transformations are also allowed). This is important because the values of $K(x|y)$ and $K(y|x)$ can be drastically different (and so do the values of $C_T(x|y)$ and $C_T(y|x)$). For example, converting the "Lena" image x to a blank image y is easy (as y can be created by a very short program), but the opposite is not. In finding the shortest program which converts one image to another, all combinations of transformations in the list must be accounted for. Since our list includes four transformations, a total of sixteen combinations are tested. Transformations are applied in the following order:

First global affine transform is used to globally align the two image. In order to achieve this goal, a six parameter affine matrix is found such that the SSIM value between the transformed source image $T_0(X)$ and the target image Y is maximized. In finding this transform, MATLAB's Genetic Algorithm Toolbox is used to find the global optimum to the objective function. The optimum matrix is then applied to the source image and its length is added to the required transform parameters in Eq. (6).

Once the images are globally aligned, a locally affine, globally smooth transform is applied to the source image. Assuming f_a and f_b are local regions of the luminance channel of the source and target images we have [23]:

$$c f_a(x, y) + b = f_b(m_1 x + m_2 y + t_x, m_3 x + m_4 y + t_y) \quad (8)$$

where m_i terms are affine parameters, and c and b are contrast and luminance change parameters. Using this registration, length of a two dimensional vector field of local geometric transformations must be included in transform parameters:

$$\overrightarrow{v}(x, y) = \begin{pmatrix} m_1 x + m_2 y + t_x - x \\ m_3 x + m_4 y + t_y - y \end{pmatrix}. \quad (9)$$

Global contrast and luminance change transform can be modelled as a linear regression problem, where α and β are selected to minimize mean squared error $\|Y - T_0(x)\|^2$. Assuming that the regression problem is formalized by $T_3(x) = L\Gamma$, we have $\Gamma = (L^T L)^{-1} L^T Y$, where:

$$\Gamma = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, L = \begin{bmatrix} (r_1 - \bar{r}) & 1 + \frac{\bar{r}}{\beta} \\ \vdots & \vdots \\ (r_N - \bar{r}) & 1 + \frac{\bar{r}}{\beta} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}. \quad (10)$$

Global Fourier power spectrum scaling is modeled in a similar way. Assuming $T_4(x) = F^{-1}\{p_1 X(\omega) + p_2\}$, and the objective is to minimize $\||Y(\omega)| - (p_1|X(\omega)| + p_2)\|^2$, we have $P = (X_\omega^T X_\omega)^{-1} X_\omega^T Y_\omega$, where:

$$P = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}, Xw = \begin{bmatrix} |x_{\omega_1}| & 1 \\ \vdots & \vdots \\ |x_{\omega_N}| & 1 \end{bmatrix}, Y_w = \begin{bmatrix} |y_{\omega_1}| \\ \vdots \\ |y_{\omega_N}| \end{bmatrix}, \quad (11)$$

and $|x_{\omega_i}|$ and $|y_{\omega_i}|$ are magnitudes of the Fourier coefficients of the source and target images, respectively.

In order to approximate the conditional Kolmogorov complexity $K(Y|X)$, all sixteen combinations of the four transformations in the list are tested. For each combination of the transformations, a transformed image $T_i(Y)$ is created. The target image X is then subtracted from the transformed source image $T_i(Y)$, resulting in a difference image with a dynamic range of $[-127 : 127]$ which contains the conditional information required to losslessly recover image Y if image X is available to the decoder. The difference image is then divided by the proposed map in Eq. (7) to remove the redundancies among neighbouring pixels and transform it into a perceptually uniform space. Finally the result is shifted by a constant in gray level and quantized into integer numbers. Figure 1 shows the process of deriving the final image which we call "Uniform image" is a perceptually compressed form of the original image Y on the condition that image X is available to the decoder.

A lossy reconstruction of the original image is possible from the uniform image. The purpose of this reconstruction is to demonstrate the visual information loss is under control. The reconstruction quality of the uniform image depends on the practical choice of parameters T_0 and C_0. Figure 2 shows reconstruction examples of the sample image Y in Figure 1 for $T_0 = 2$. It is evident that as the choice of the parameter C_0 can greatly affect the visual quality of the reconstructed image, and as C_0 increases the SSIM between the reconstructed image and the original image increases. The optimum values for these parameters are tuned using small datasets. In this paper, we select $T_0 = 2$ and $C_0 = 0.1$ in our simulations.

IV. Applications

Here we demonstrate NCCD through real-world image classification and recognition problems. It is worth noting that NCCD is applied as a generic image similarity measure. Unlike many other image classification and recognition methods, the NCCD measure does not require any domain adaptation, and does not need any training using application specific samples. In this section, NCCD's performance is tested against the domain-specific texture classification and face recognition datasets.

A. Texture Classification

We apply NCCD to a variety of texture datasets commonly used in literature and compare our results to those of two

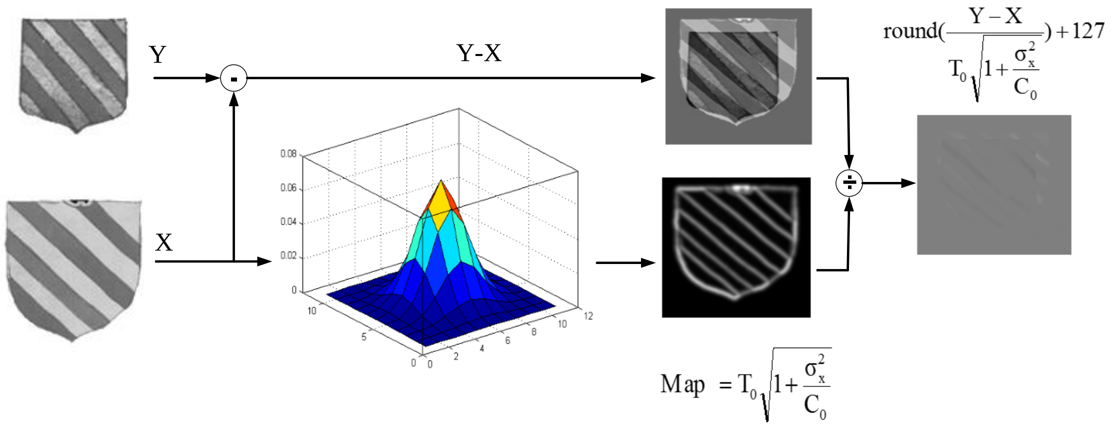
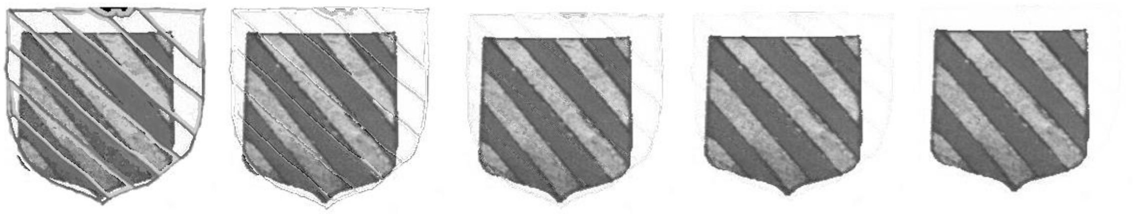

Y
Y-X
X
$\mathrm{round}(\frac{Y-X}{T_0\sqrt{1+\frac{\sigma_x^2}{C_0}}})+127$
0.08
0.06
0.04
0.02
0
10
5
0
0
2
4
6
8
10
12
$\mathrm{Map} = T_0\sqrt{1+\frac{\sigma_x^2}{C_0}}$

Fig. 1: Deriving the uniform image.



$C_0=\frac{1}{50}$,SSIM=0.65 $C_0=\frac{1}{5}$,SSIM=0.77 C_0=1,SSIM=0.87 C_0=5,SSIM=0.94 C_0=50,SSIM=0.99

Fig. 2: Reconstruction of the encoded source image for various C_0 values.

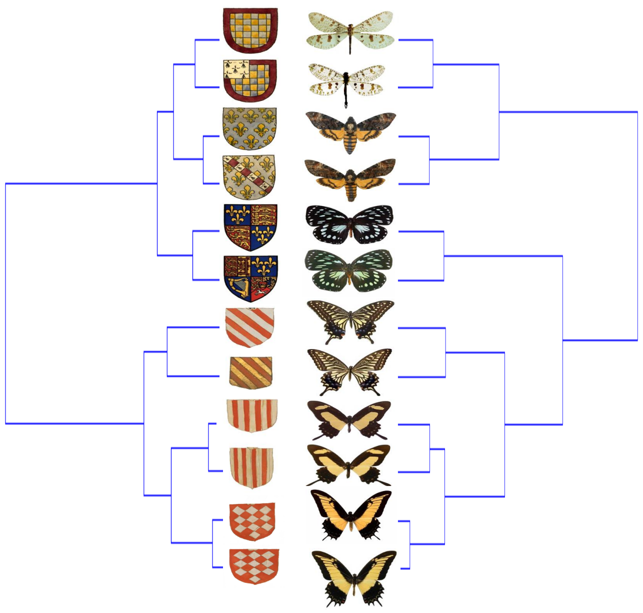


Fig. 3: Texture clustering using NCCD.

compression based distance methods. In the first step, we classify two small datasets of images of various sizes and colors. Figure 3 shows the result of clustering the *Heraldic Shields* and the *Butterflies* [12] dataset. Both datasets contain 12 images of different sizes, which are hierarchically clustered according to their similarity in texture using the linkage method. It can be observed that the clustering is consistent with human intuition in both scenarios. In order to compare the performance of NCCD with other compression based distance methods, we perform leave-one-out classification experiments on a number of datasets used in [12]. The classification reuslts on these datasets are then compared to those of CK-1 distance [12], which uses a compression distance based on MPEG-I video encoder to capture texture similarity, and a compression distance called Sparsity Based Compression which uses sparsity as a measure of compression, and a sparse representation-based approach to encode the information content of an image using information of another image introduced in [13]. A brief introduction of these datasets is presented in the following.

Tire Treads is a collection of tire tracks, and contains 48 images of 3 tires rolled in 16 different directions [12].

Brodatz Texture is a collection of 1, 792 man-made and natural texture images digitalized from a reference photographic album for designers [24].

Camouflage is a collection of 80 random orientations of 9 modern US military camouflage [12].

VTT Wood is a collection of 200 images of wood which are classified into two subset of healthy and defective woods with 40 types of wood defects [25].

VisTex is a collection of homogeneous texture images and texture scenes created by MIT Vision and Texture Group, which do not conform to rigid frontal plane perspectives and studio lighting conditions [26].

In leave-one-out cross validation scheme, a query image is selected from the dataset and all images in the dataset are ranked based on their distance to the query image in ascending order. The first K images are used to develop a hypothesis about the type of the query image. The hypothesis is then checked, and the performance of the classification method is defined as the ratio of correctly classified images to the total number of the images in the dataset. In each case the first image in the results is the same as the query image, and the following images are the closest images in distance to the query image. The results for classification of the above datasets using NCD, NCCD, CK-1 as well as Sparsity Based Compression (SBC) distance method proposed in [13], are provided in Table I, where we use leave-one-out scheme and 1-Nearest Neighbor framework in order to create comparable results to those provided in [13]. It can be observed that the performance of NCCD is comparable to, or better than existing state-of-the-art compression based classification methods.

B. Face Recognition

We also test the proposed NCCD framework for face recognition on two of the widely cited databases in the literature. Many algorithms and databases have been developed which report success in recognizing faces under different illumination, pose and occlusion conditions.

We apply the distance to AT&T [27] and Yale [28] face datasets, and compare our results to those of CK-1 [12] and SBC [13] compression distances. AT&T [27] dataset consists of images of 40 individuals in 10 different poses, taken under different illumination conditions and facial expressions and details. Yale face dataset [28] contains 165 grayscale images in GIF format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink. Figure 4 shows examples of image retrieval in AT&T dataset. To create this figure, the NCCD among the query image and all images in the dataset is computed, and images in the dataset are ranked according to their mutual NCCD score with the query image in ascending order. Note that in cases where NCCD fails to find a match of the same face in a different position, the retrieved image has a visual resemblance to the query image. For instance, if the person in the query image is wearing glasses, the first ten faces retrieved from the dataset are also wearing glasses. It is evident that NCCD is capable of distinguishing individual faces with different facial expressions in the presence of occlusions such as glasses. Similar to [13], we test the NCCD framework by measuring its clustering performance on AT&T and Yale face datasets. For each image in the dataset, a distance vector is created based on NCCD among the image and all other images in the dataset. These vectors are then inserted into the rows of a $N \times N$ NCCD matrix, and the matrix is used by a standard spectral clustering algorithm [30] to create a vector of cluster labels. The labels are then used to predict the accuracy of the framework using Hungarian algorithm [31]. Table II shows the results of clustering AT&T and Yale face datasets using NCCD compared to CK-1 and SBC as reported in [13].

V. Computational Complexity

The proposed implementation of the NCCD framework requires approximately 20 seconds to find the distance among a pair of 128×128 images on a core-i3 Intel processor. Most of this time is taken by computing the sixteen transformations, including global and local affine transforms. Table III compares computational complexity of the NCCD framework with those of CK-1 [12], SBC [13], and SSIM [3], where computation time of the algorithms are measured by the same pair of 128 × 128 images.

VI. Ablation Test

In order to quantify the effects of the DNT and each of the transformation in perceptual NCCD framework we devise an ablation test, where each of these features are removed from the framework, and the performance is calculated on all datasets without that particular feature. Table IV shows the performance of NCCD framework with each of the features removed, where NCCD-1 is perceptual NCCD framework without the Divisive Normalization Transform (DNT), NCCD-2 is the framework without Global contrast and luminance change transformation, NCCD-3 is the framework without Global Fourier power spectrum scaling, NCCD-4 is the framework without Global affine transform, and NCCD-5 is the framework without Local registration transformation.

It can be inferred from IV that the DNT has a key role in performance of the NCCD framework. It can also be observed that each of the other transformations play an important role for applicable scenarios. For example, while the Global contrast and luminance change transform play an important role in the performance of the framework in Tire tracks [12], and VVT Wood [25] datasets where many of the images are rotated versions of each other, it has less importance in the performance of the framework for face recognition datasets such as AT&T [27] and Yale [28]. It is also notable that in such datasets, transformations such as Local registration transformation, and Global Fourier power spectrum scaling play a more important role.

VII. Conclusion

This work aims to develop a generic perceptual image similarity measure based upon the theoretic groundwork of Kolmogorov complexity and the NID metric. The most important contribution of this paper is to propose a practical framework of NCCD for the approximation of NID, and use Divisive Normalization Transform (DNT) as a nonlinear model to remove statistical redundancies among natural images in finding perceptually relevant information content of the image. The proposed framework is flexible and expandable to include

TABLE I: Classification performance of various datasets using NCD [29], NCCD, CK-1 [12], and SBC [13]

Dataset	NCD [29] (%)	NCCD (%)	CK-1 (%) [12]	SBC(%) [13]
Brodatz [24]	1.2	73.2	54.0	**76.2**
Camouflage [12]	1.6	81.9	**87.5**	87.0
Tire tracks [12]	0.8	**93.8**	79.2	79.2
VTT Wood [25]	0.9	**91.9**	80.5	85.2
VisTex [26]	1.3	**39.3**	32.9	N/A

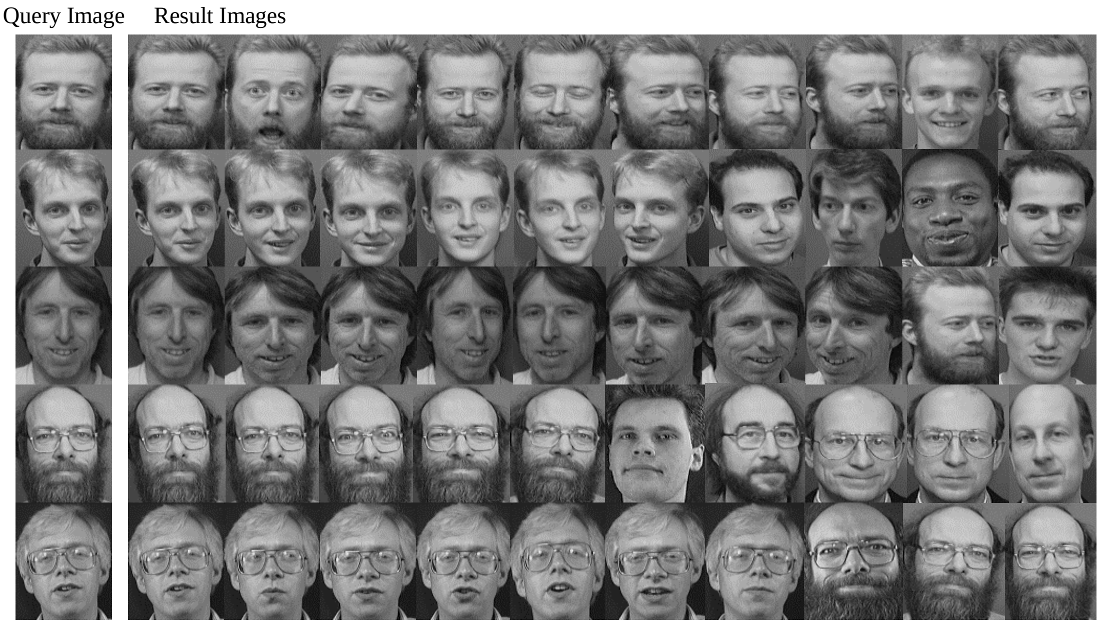


Fig. 4: Image retrieval in AT&T Face Database [27]. For each row, the left-most is the query image, and the rest are retrieved images of the least NCCD distance in ascending order.

TABLE II: Clustering performance of AT&T and Yale face datasets using NCD [29], NCCD, CK-1 [12], and SBC [13]

Dataset	NCD [29] (%)	NCCD (%)	CK-1 (%) [12]	SBC(%) [13]
AT&T [27]	10.1	**82.8**	76.5	81.6
Yale [28]	8.3	**73.1**	64.1	65.9

TABLE III: Computational complexity of NCCD, CK-1 [12], SBC [13], and SSIM [3] for a pair of 128 $\times$ 128 images

Algorithm	NCCD	CK-1 [12]	SBC [13]	SSIM [3]
Time (Seconds)	20.6	8.1	11.9	0.8

TABLE IV: Ablation test

Dataset	NCCD-1 (%)	NCCD-2 (%)	NCCD-3 (%)	NCCD-4 (%)	NCCD-5 (%)
Brodatz [24]	24.2	54.3	49.2	18.9	53.1
Camouflage [12]	17.3	60.9	62.1	14.3	59.8
Tire tracks [12]	21.9	52.1	40.0	20.4	65.9
VTT Wood [25]	20.1	63.8	65.9	16.5	67.9
VisTex [26]	19.1	70.2	59.0	15.1	69.1
AT&T [27]	14.7	32.8	42.2	60.1	33.1
Yale [28]	18.2	28.9	48.9	64.0	40.2

any image transformations that may help find the shortest description that converts one image to another and vice versa. Experimental results show that the proposed method, without domain adaptation and without training, is competitive against state-of-the-art compression and sparsity based image distance measures in several texture classification and face recognition tests.

Acknowledgment

Authors would like to thank NSERC, SOSCIP, and Ontario Centres of Excellence for their continued support and funding of this project.

References

[1] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. Morgan & Claypool Publishers, Mar. 2006.

[2] Z. Wang and A. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *Signal Processing Magazine, IEEE*, vol. 26, pp. 98 –117, jan. 2009.

[3] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[4] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Transactions on Image Processing*, vol. 18, pp. 2385–2401, Nov 2009.

[5] and E. P. Simoncelli, "An adaptive linear system framework for image distortion analysis," in *IEEE International Conference on Image Processing 2005*, vol. 3, pp. III–1160, Sep. 2005.

[6] and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 2, pp. ii/573–ii/576 Vol. 2, March 2005.

[7] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi, "The similarity metric," *IEEE Trans. Info. Theory*, vol. 50, pp. 3250–3264, Dec. 2004.

[8] H. Barlow, *Possible principles underlying the transformation of sensory messages*. Cambridge, MA: MIT Press, 1961.

[9] S. Lyu, "Divisive normalization: Justification and effectiveness as efficient coding transform," in *In Advances in Neural Information Processing Systems 23*, pp. 1522–1530, 2010.

[10] J. Chua and P. Tischer, "Focusing the normalised information distance on the relevant information content for image similarity," in *Digital Image Computing: Techniques and Applications (DICTA), 2010 International Conference on*, pp. 1–7, 2010.

[11] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual review of neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.

[12] B. J. L. Campana and E. J. Keogh, "A compression based distance measure for texture," *Stat. Anal. Data Min.*, vol. 3, pp. 381–398, Dec. 2010.

[13] T. Guha and R. K. Ward, "On image similarity, sparse representation and kolmogorov complexity," *CoRR*, vol. abs/1206.2627, 2012.

[14] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*. Berlin, 2nd edition: Springer, 1997.

[15] X. Wu and N. Memon, "Context-based, adaptive, lossless image codec," *IEEE Trans. Comm*, vol. 45, pp. 437–444, Apr. 1997.

[16] N. Tran, "The normalized compression distance and image distinguishability," in *The 19th IS&T/SPIE Symposium on Electronic Imaging Science and Technology*, (San Jose), Jan. 2007.

[17] N. Nikvand and Z. Wang, "Generic image similarity based on kolmogorov complexity," in *Proceedings of International Conference on Image Processing (ICIP) 2010*, 2010.

[18] Y. Karklin and E. P. Simoncelli, "Efficient coding of natural images with a population of noisy linear-nonlinear neurons," in *In Adv. Neural Information Processing Systems (NIPS*11*, pp. 13–15, MIT Press, 2012.

[19] M. Carandini and D. J. Heeger, "Normalization as a canonical neural computation," *Nat. Rev. Neurosci.*, vol. 13, no. 1, pp. 51–62, 2012.

[20] S. Lyu and E. Simoncelli, "Nonlinear image representation using divisive normalization," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, 2008.

[21] S. Deneve, A. Pouget, and P. Latham, "Divisive normalization, line attractor networks and ideal observers," in *Advances in Neural Processing Systems*, pp. 104–110, The MIT Press, 1999.

[22] K. Louie, M. Khaw, and P. Glimcher, "Normalization is a general neural mechanism for context-dependent decision making," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 6139–6144, 2013.

[23] S. Periaswamy and H. Farid, "Elastic registration in the presence of intensity variations," *IEEE Trans. Medical Imaging*, vol. 22, pp. 865–874, July 2003.

[24] T. Randen, "Brodatz texture image database," http://www.ux.uis.no/~tranden/brodatz.html.

[25] O. Silvén, M. Niskanen, and H. Kauppinen, "Wood inspection with non-supervised clustering," *Mach. Vision Appl.*, vol. 13, pp. 275–285, Mar. 2003.

[26] MIT Vision and Modeling Group, "Vision Texture Database," http://vismod.media.mit.edu/vismod.

[27] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, pp. 138–142, 1994.

[28] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 711–720, 1997.

[29] R. Cilibrasi and P. M.B. Vitányi, "Clustering by compression," *IEEE Trans. Info. Theory*, vol. 51, pp. 1523–1545, Apr. 2005.

[30] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. NIPS.*, pp. 849–856, MIT Press, 2001.

[31] C. Papadimitriou and K. Steiglitz, *Combinatiorial Optimization: Algorithms and Complexity*. Dover Publications, 1998.