

Manuscript version: Author's Accepted Manuscript

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:

<http://wrap.warwick.ac.uk/112087>

How to cite:

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

Image Similarity Using Sparse Representation and Compression Distance

Tanaya Guha, *Student Member, IEEE*, and Rabab K Ward, *Fellow, IEEE*

Abstract

A new line of research uses compression methods to measure the similarity between signals. Two signals are considered similar if one can be compressed significantly when the information of the other is known. The existing compression-based similarity methods, although successful in the discrete one dimensional domain, do not work well in the context of images. This paper proposes a sparse representation-based approach to encode the information content of an image using information from the other image, and uses the compactness (sparsity) of the representation as a measure of its compressibility (how much can the image be compressed) with respect to the other image. The more sparse the representation of an image, the better it can be compressed and the more it is similar to the other image. The efficacy of the proposed measure is demonstrated through the high accuracies achieved in image clustering, retrieval and classification.

Index Terms

Image similarity, Compression, Kolmogorov complexity, Overcomplete dictionary, Sparse representation

I. INTRODUCTION

Measuring the similarity between a pair of images is of critical importance to many multimedia information processing systems involving retrieval, enhancement, copy detection, quality assessment, clustering and classification. Given the long history of image similarity evaluation, the volume of literature

The authors are with the Image and Signal Processing Laboratory, department of Electrical and Computer Engineering, the University of British Columbia, Vancouver, BC, Canada.

e-mail: tanaya@ece.ubc.ca, rababw@ece.ubc.ca

on this topic is large and diverse. Widely used similarity measures such as the Euclidean distance, the Mean Squared Error and other norm-based measures work well in specific cases, but they are often criticized for not corresponding well with our visual perception of similarity [1]. Another popular approach to describe the visual content of images is to extract a set of meaningful features. The similarity between two images is then computed in terms of the similarity between their features. However, the success of this approach is limited by the availability, selection and extraction of a good set of meaningful features, demanding specific knowledge of the application and the data.

Recently, there has been an interest in developing image similarity measures using *compression* methods [2]–[6]. In this approach, two signals are considered similar if one can be compressed significantly when the information of the other is provided. The advantages of these methods are that they are *parameter-free* (the only choice the user has to make is which compression algorithm to use) and *generic* (they assume no prior knowledge of the application, and can be applied, without modification, to a variety of problems).

The compression-based similarity methods rely on a new mathematical theory of similarity which is in turn based on the idea of the *Kolmogorov complexity* [2], [3]. The Kolmogorov complexity (also known as the algorithmic entropy) is a theoretical measure of randomness of a given data, and in general, is a non-computable quantity. In practice, it is often approximated by the *length of the compressed data*. Intuitively, the more a given data can be compressed, the lower is its complexity.

The compression-based similarity measures have been shown to be highly effective in clustering and classifying discrete, uni-dimensional data such as text and protein sequences [2], [3]; but their successful application in the context of real-valued, higher-dimensional data like images is scarce. For effectively measuring the similarity between two signals, the compressor being employed needs to satisfy certain properties so as to be a *normal* compressor [3]. However, most state-of-the-art compressors for images (such as JPEG, JPEG2000) are *not* normal, and the normal compressors (compressors of the Lempel-Ziv family) do not work well on images [5]. Existing methods [7]–[9] transform images into strings in order to take advantage of the normal data compressors, and thus lose the important spatial information. Another serious obstacle lies in evaluating and approximating the *conditional compression* (a quantity that measures how much can a given data be compressed w.r.t. another data) which is the key component in every compression-based similarity measure.

In this paper, we propose a sparse representation-based approach to encode the information content of an image; and use the compactness (sparsity) of the representation of the image as a measure of its compressibility i.e. how much can the image be compressed. The more sparse the representation of an

image, the better it can be compressed.

In order to design a similarity measure that correlates well with the human perception, we learn a set of basis elements (collectively called a *dictionary*) from the images. This approach empowers us to build a *cortex-like representation* of an image. In 1996, Olshausen and Field have shown that the basis elements that resemble the properties of the receptive field of simple cells in the primary visual cortex can be learnt from input images [10]. The keys to building such a cortex-like dictionary are: (i) a *sparsity prior* - an assumption that it is possible to describe the input image using a small number of basis elements, and (ii) *overcompleteness* - the number of basis elements in the dictionary is greater than the vector space spanned by the input vectors. Given a pair of images, our method learns a dictionary for each image and computes how sparsely can one image be approximated using the dictionary extracted from the other, with a required precision.

The rest of the paper is organized as follows. Section II briefly describes the related work on compression-based distances, section III proposes the sparse representation-based distance measure, and section IV presents experimental results. Section V concludes the article with possible directions to future work.

II. PREVIOUS WORK

The work of Kolmogorov and others [11]–[13] on how to measure data complexity has been influential in many areas of knowledge, across multiple disciplines. The notion of complexity of a string is related to its randomness. For example, the binary string 1101010001 is considered more complex compared to the string 0101010101, because the latter contains a regularity (repeating pattern) and therefore is less random. Kolmogorov complexity formalizes this concept.

Given a finite object, such as a binary string \mathbf{x} , its *Kolmogorov complexity* $K(\mathbf{x})$ is defined as the length of the shortest program that can effectively produce \mathbf{x} on a universal computer, such as a Turing machine [14]. The Kolmogorov complexity, however, is non-computable in general. In practice, it is often approximated by the length or the file size of the compressed data. Intuitively, the more a given data can be compressed, the lower is its complexity.

A. Compression-based distance measures

Recently, Kolmogorov's theory of complexity has been used to address the problem of similarity measurement [2], [3]. Given two signals \mathbf{x} and \mathbf{y} , a distance metric, known as the *Normalized Information Distance* (NID) is developed using $K(\mathbf{x})$ and the conditional Kolmogorov complexity $K(\mathbf{x}|\mathbf{y})$. $K(\mathbf{x}|\mathbf{y})$

is defined as the length of the shortest program used by a universal computer to generate \mathbf{x} when \mathbf{y} is known.

Due to the non-computable nature of the Kolmogorov complexity, a practical analog of the NID metric is proposed based on standard compression methods. This is called the *Normalized Compression Distance* (NCD). Intuitively, NCD considers \mathbf{x} and \mathbf{y} to be similar if one can be significantly compressed when the information of the other is provided. It is defined as follows:

$$\text{NCD}(\mathbf{x}, \mathbf{y}) = \frac{\max\{C(\mathbf{x}|\mathbf{y}), C(\mathbf{y}|\mathbf{x})\}}{\max\{C(\mathbf{x}), C(\mathbf{y})\}} \quad (1)$$

The conditional compression $C(\mathbf{x}|\mathbf{y})$ is approximated as follows:

$$C(\mathbf{x}|\mathbf{y}) = C(\mathbf{xy}) - C(\mathbf{y}) \quad (2)$$

where $C(\mathbf{xy})$ denotes the compressed length of the concatenation of \mathbf{x} and \mathbf{y} .

The NCD metric has been shown to be effective in clustering mitochondrial genomes, languages and music [3]. Following the success of NCD, different versions of compression-based distance measures have been proposed; for example, a *Compression-based Dissimilarity Measure* (CDM) is proposed in the context of parameter-free data mining and is shown to be useful for anomaly detection, clustering and classification of text, DNA and time-series data [15]. CDM is defined as

$$\text{CDM}(\mathbf{x}, \mathbf{y}) = \frac{C(\mathbf{xy})}{C(\mathbf{x}) + C(\mathbf{y})} \quad (3)$$

Other applications of compression-based distances include symbolic music clustering [16] and plagiarism detection [17]. The idea of compression, independent from NCD, has also been used to design a pattern representation scheme for automatic categorization of music, voice, genome, etc. [4]; but this method requires encoding media data input into text.

B. Compression-based distances for images

In the context of images, however, successful application of the compression-based distance measures is scarce. We identify two major reasons behind that.

- The success of the compression-based distances heavily depends on the availability of a *normal* compressor. A compressor is normal only if it satisfies certain conditions such as idempotency, monotonicity, symmetry, etc. (please refer to [3] for details). The problem is that most state-of-the-art image compressors (such as JPEG, JPEG2000) are *not* normal, and the normal compressors (such as the compressors of the Lempel-Ziv family) do not work well on images [5].

- Another serious obstacle lies in evaluating and approximating the conditional complexity terms such as $C(\mathbf{x}|\mathbf{y})$ in NCD. These terms are the key components in a compression-based measure. The existing compression-based methods (whether or not they involve images) either approximate the conditional compression $C(\mathbf{x}|\mathbf{y})$ with $C(\mathbf{xy}) - C(\mathbf{y})$ or use a simplified definition so as not to include any conditional term (as in (3)). Direct evaluation of $C(\mathbf{x}|\mathbf{y})$ is usually bypassed mainly to retain the simplicity of the compression-based measures since evaluating $C(\mathbf{x}|\mathbf{y})$ accurately requires delving into the complicated standards and algorithms of data or image compression. This also makes the compression-based methods difficult to improve upon.

Clearly, the straightforward extension of the methods that work perfectly well on discrete, one-dimensional data has not been very promising in the context of images. In the pursuit of alternatives, a new image encoder is proposed based on the finite context model and preliminary results on a face database are provided [5]. Another recent approach, namely the CK-1 method, uses the MPEG1 video compressor to measure image similarity [6]. This method takes advantage of the temporal redundancy reduction step in video compression which performs inter-frame block matching. In this approach, a two-frame video consisting of the images to be compared is created. One frame is compressed with reference to the other frame using a standard video compressor. The compressed file size of the video is used to approximate the closeness between the pair of images. This method has been shown to be useful in texture classification.

III. THE PROPOSED APPROACH

A natural way of measuring the similarity between two given images is to quantify how well either image can be represented using the information of the other. The more similar the images, the better is the representation of one image in terms of the other. Our method formalizes this intuitive idea of similarity using a sparse representation-based approach.

A. Sparsity as a measure of data complexity

It is well-known that sparsity of representation plays a key role in achieving good compression. For example, the superiority of JPEG2000 is mainly attributed to the capability of the wavelet transform toward representing an image more sparsely than the DCT used in JPEG. Intuitively, the more sparse the representation of a signal is, the fewer are the components needed to capture the signal's information content and the better it can be compressed.

Sparsity thus can be seen as a direct measure of the randomness or complexity of the data. A natural image usually exhibits many repeated structures which can be discovered through its decomposition over a set of properly chosen basis functions. Due to the presence of redundancy, only a few basis functions are required to capture the significant information content of such images, resulting in a sparse representation. In the case where such structures are rare (e.g. in random Gaussian noise), there is no way to represent the data using a small number of basis elements. This indicates that as the complexity of a signal increases, more and more components are needed to represent the signal with a desired accuracy i.e. its sparsity decreases in the transform domain. This inherent connection between sparsity and data complexity is exploited in our proposed distance measure.

B. Sparse Representation-based distance measure

The basic idea in sparse signal analysis is to represent a signal by a linear combination of a small number of basis functions. Consider a signal $\mathbf{b} \in \mathbb{R}^m$ represented as a linear combination of n basis functions or atoms,

$$\mathbf{b} = \mathcal{D}\mathbf{a} \quad (4)$$

where the dictionary $\mathcal{D} \in \mathbb{R}^{m \times n}$ and its columns are the basis functions or atoms. If the values of the majority of components in $\mathbf{a} \in \mathbb{R}^n$ are 0 (or close to 0), we say that \mathbf{x} has a *sparse* representation w.r.t. \mathcal{D} . For orthogonal bases like Fourier, \mathcal{D} is a square matrix i.e. $m = n$. For those cases where the number of basis vectors is greater than the dimensionality of the input signal i.e. where $m < n$, \mathcal{D} is said to be *overcomplete*. An overcomplete dictionary offers greater flexibility in representing the essential structures in a signal, which in turn leads to higher sparsity in the transform domain. Such representation also has advantages such as robustness to additive noise and occlusion [18].

1) *learning the dictionaries*: Let us consider an image X . A set of k random, possibly overlapping patches (each of dimension $\sqrt{m} \times \sqrt{m}$) is extracted from X . Every patch is converted to a vector of length m and the patches are concatenated to form a matrix $\mathbf{B}_x \in \mathbb{R}^{m \times k}$. In order to build a perceptually meaningful model for X , we intend to learn an overcomplete dictionary $\mathcal{D}_x \in \mathbb{R}^{m \times n}$ that has n atoms ($m < n$) using the local patches in \mathbf{B}_x as input. However, greater difficulties arise with a set of overcomplete bases. An overcomplete dictionary matrix creates an underdetermined system of linear equations having an infinite number of solutions. Knowing that the natural signals are sparsely representable, often in such cases, we seek the sparsest solution i.e. we want the vector \mathbf{a} to contain as few non-zero elements as possible.

Our objective is to learn \mathcal{D}_x such that each patch (column) $\mathbf{b}_{x_i} \in \mathbf{B}_x$ can be closely approximated as a linear superposition of a small number of atoms in \mathcal{D}_x . This is achieved by solving the following sparse optimization problem:

$$\min_{\{\mathcal{D}_x, \mathbf{a}_x\}} \sum_i \|\mathbf{a}_{x_i}\|_p \quad \text{s.t.} \quad \forall i, \|\mathbf{b}_{x_i} - \mathcal{D}_x \mathbf{a}_{x_i}\|_2 \leq \epsilon \quad (5)$$

where the vector $\mathbf{a}_{x_i} \in \mathbb{R}^n$ is the sparse representation of the patch $\mathbf{b}_{x_i} \in \mathbb{R}^n$. The sparse representation of \mathbf{B}_x w.r.t. \mathcal{D}_x is denoted as the matrix $\mathbf{A}_x = [\mathbf{a}_{x_1} | \mathbf{a}_{x_2} | \dots | \mathbf{a}_{x_k}]$. The value of p is typically 0 or 1 and ϵ denotes the reconstruction error controlled by the user.

Note that, with $p = 0$ (the ℓ_0 seminorm that counts the number of non-zero elements in a vector) equation (5) becomes non-convex, and solving it exactly is an NP hard problem. Approximate solution is found instead using either greedy algorithms [19] or using convex relaxation [20]. The convex relaxation methods use $p = 1$ (the ℓ_1 norm) to transform (5) into a convex problem.

We employ a fast dictionary learning algorithm called the K-SVD algorithm [21] which provides an approximate solution to (5) for the ℓ_0 case. It performs two steps at every iteration: (i) sparse coding and (ii) dictionary update. In the first step, the dictionary \mathcal{D}_x is fixed and \mathbf{a}_{x_i} is computed by a greedy algorithm called *Orthogonal Matching Pursuit* (OMP) [19]. Next, the atoms of \mathcal{D}_x are updated sequentially, allowing the relevant coefficients in \mathbf{a}_x to change as well. For the details of this algorithm, please refer to the original K-SVD paper [21].

2) *Sparse representation-based complexity functions*: We define two quantities that measure the compressibility (how much can an image be compressed) of an image by (i) using its own dictionary, and (ii) using the dictionary extracted from the other image, Y . We name these terms as the *Sparse complexity* and the *Relative sparse complexity*, respectively.

Definition 1. Given an image X , its *Sparse Complexity* $S(X, \mathcal{D}_x)$ is defined as the sparsity of \mathbf{A}_x averaged over the number of columns in \mathbf{A}_x i.e.

$$S(X, \mathcal{D}_x) = \frac{1}{k} \|\mathbf{A}_x\|_p = \frac{1}{k} \sum_{i=1}^k \|\mathbf{a}_{x_i}\|_p \quad (6)$$

Therefore, for $p = 0$, $S(X, \mathcal{D}_x)$ is the average number of non-zero coefficients required to reconstruct a column of \mathbf{B}_x using \mathcal{D}_x , up to a required precision ϵ . Smaller value of $S(X, \mathcal{D}_x)$ indicates higher compressibility (lower complexity) of X .

Properties of $S(X, \mathcal{D}_x)$:

- $S(X, \mathcal{D}_x) > 0$ for non-empty X , and is equal to 0 otherwise.
- Considering that X is represented by \mathbf{A}_x and hence XX is represented by $[\mathbf{A}_x | \mathbf{A}_x]$, we have $S(XX, \mathcal{D}_x) = S(X, \mathcal{D}_x)$.

This property (idempotency) follows from the averaging operation and indicates that the sparse complexity function can compress the duplicate entries.

Given another image Y , the compression-based measures attempts to approximate how much can the image X be compressed when additional information about Y is available. As discussed before, this conditional quantity, is difficult to approximate and that limits the success of these measures. We hence define a slightly different complexity term that measures *how much information about X is contained in Y* . We name this term as the *Relative Sparse Complexity*, .

Let $\mathcal{D}_y \in \mathbb{R}^{m \times n}$ be the dictionary pertaining to the image Y learnt in the same manner as \mathcal{D}_x (refer to(5)). The image X can be approximated in terms of the dictionary of Y as follows:

$$\min_{\mathbf{a}_{x|y}} \sum_{i=1}^k \|\mathbf{a}_{x|y_i}\|_p \quad \text{s.t.} \quad \|\mathbf{b}_{x_i} - \mathcal{D}_y \mathbf{a}_{x|y_i}\|_2 \leq \epsilon \quad (7)$$

where $\mathbf{a}_{x|y_i} \in \mathbb{R}^n$ is the sparse representation of \mathbf{b}_{x_i} w.r.t. \mathcal{D}_y and $\mathbf{A}_{x|y} = [\mathbf{a}_{x|y_1} | \mathbf{a}_{x|y_2} | \dots | \mathbf{a}_{x|y_k}]$ is the sparse representation of \mathbf{B}_x w.r.t. \mathcal{D}_y .

Definition 2. Given two images X and Y , the Relative Sparse Complexity $S(X, \mathcal{D}_y)$ is defined as the sparsity of $\mathbf{A}_{x|y}$ averaged over the number of columns in $\mathbf{A}_{x|y}$.

$$S(X, \mathcal{D}_y) = \frac{1}{k} \|\mathbf{A}_{x|y}\|_p = \frac{1}{k} \sum_{i=1}^k \|\mathbf{a}_{x|y_i}\|_p \quad (8)$$

Therefore, for $p = 0$, $S(X, \mathcal{D}_y)$ becomes the average number of non-zero coefficients required to reconstruct a column of \mathbf{B}_x using \mathcal{D}_y , up to a required precision ϵ . A smaller value of $S(X, \mathcal{D}_y)$ indicates that X is efficiently represented by the information extracted from Y i.e. X and Y have higher similarity.

Properties of $S(X, \mathcal{D}_y)$:

- $S(X, \mathcal{D}_y) > 0$ for non-empty Y , and 0 otherwise.
- $S(XY, \mathcal{D}_y) = S(YX, \mathcal{D}_y)$ (symmetry)
- $S(X, \mathcal{D}_y) > S(X, \mathcal{D}_x)$ for $X \neq Y$. This is because, in general, X is expected to be more efficiently (sparsely) approximated using \mathcal{D}_x - the dictionary trained on itself, than \mathcal{D}_y - a dictionary trained on a different image.

3) *The distance measure:* Based on the two terms defined above, a sparse representation-based distance measure d_S is defined as follows:

$$d_S(X, Y) = \frac{S(X, \mathcal{D}_y) + S(Y, \mathcal{D}_x)}{S(X, \mathcal{D}_x) + S(Y, \mathcal{D}_y)} - 1 \quad (9)$$

The proposed form of d_S is much similar to that of the compression-based CK-1 distance measure [6]. From the property of the relative sparse complexity we have

$$S(X, \mathcal{D}_y) > S(X, \mathcal{D}_x) \quad \text{and} \quad S(Y, \mathcal{D}_x) > S(Y, \mathcal{D}_y)$$

Hence,

$$\frac{S(X, \mathcal{D}_y) + S(Y, \mathcal{D}_x)}{S(X, \mathcal{D}_x) + S(Y, \mathcal{D}_y)} > 1 \quad \text{for} \quad X \neq Y.$$

Intuitively, d_S measures how efficient, on average, is it to approximate one image X using the information of Y extracted in the form of a dictionary of its dominant local structures. The smaller the values of d_S the higher is similarity between the two images.

Properties of d_S :

- *Non-negativity:* d_S is always non-negative, the lowest value of d_S is 0 when $X = Y$.
- *Symmetry:* Clearly, d_S is symmetric i.e. $d_S(X, Y) = d_S(Y, X)$. Symmetry is an important property for a similarity or dissimilarity measure because many algorithms (e.g. spectral clustering) rely on this property.
- *Metricity:* d_S does not follow the metric axiom of triangle inequality and hence cannot be called a metric. It would have been mathematically convenient if d_S was a metric. However, many researchers have argued that perceptual distances are typically non-metric in nature [22], [23].

Note that, we have used $p = 0$ to compute the complexity functions because our dictionary learning method uses greedy ℓ_0 approximation. If ℓ_1 optimization is used to learn the dictionaries, it would be better to use $p = 1$ for the definitions.

IV. EXPERIMENTAL VALIDATION

In order to establish the generality of the proposed distance measure, we perform experiments on a variety of applications. We first perform experiments to evaluate the compatibility of the proposed measure with the human perception of similarity. This is followed by clustering, retrieval and classification experiments involving larger datasets. The datasets that we choose contain real-world images from different domains like biology, biometrics, medicine and natural textures.

A. Implementation Details

Practically, there are 4 parameters to be set: the patch size (\sqrt{m}), the number of patches to be extracted from each image (k), the number of dictionary elements (n) and the reconstruction error (ϵ). Unfortunately, there is no theoretical guidelines to determine the values of these parameter, so we rely on previous work and empirical methods. We have used the same parameter values for all experiments, unless mentioned otherwise. Below, we describe how the parameter values are chosen for this particular work.

Patch size (\sqrt{m}) and automatic scale selection: The patch size determines the spatial scale at which an image is analyzed. For simplicity and speed, we analyze each image at a single scale, but use a simple technique to *automatically* select the (sub)optimal scale. A 2D Laplacian of Gaussian (LoG) filter is applied to each image to detect the local maxima points (keypoints) at four different scales. The scale at which the maximum number of keypoints are detected is chosen as the (sub)optimal scale for that image. The image is downsampled accordingly and a set of patches are extracted. For example, if the scale is found to be 2, the image is downsampled by a factor of 2 and then patches of size 8×8 i.e. $\sqrt{m} = 8$ are extracted. This particular patch size is chosen in order to be consistent with most of the compression based algorithms (e.g. JPEG1) which process 8×8 blocks. The automatic scale selection is performed on all images for all datasets except for the VVT Wood dataset due to the small dimensions (64×64) of the original images.

Number of patches (k): In order to train a dictionary, a large number of patches need to be extracted. The color images are first converted to grayscale to achieve *color invariance*. It is also important that the randomly extracted patches contain important structural information of the image and do not come from the homogeneous regions of the image only. This is accomplished by selecting the patches whose energy levels are above an empirically set threshold. A collection of $k = 3000$ such patches are extracted from every image and is used to train its corresponding dictionary. The input patches for dictionary learning have zero mean and unit standard deviation which account for *luminance and contrast invariance*.

Overcompleteness (n/m): Since we intend to learn an overcomplete dictionary, we must have $n > m$. The ratio n/m is called the *overcompleteness factor*. It has been shown that for small overcompleteness factor, sparse representation is stable in the presence of noise [24]. Thus we set $n/m = 2$, where $m = 64$.

Reconstruction error (ϵ): We used $\epsilon = 0.1$ which means that the input vector is reconstructed with at least 90% accuracy. Note that a lower reconstruction error can produce a better dictionary, but requires more computation and more importantly, may cause overfitting.

B. Correlation with Human Perception

It is important that the distance measure between images correlate with human perception. We begin with measuring the similarities between a reference image (Fig. 1(a)) and its distorted versions (Fig. 1(b)-(e)) as well as a completely unrelated image (Fig. 1(f)). We also compare our results with PSNR and the well-known *Visual Information Fidelity* (VIF) [25] (values closer to zero indicates lower similarity) similarity measure. Figure 1 shows that the proposed measure correlates well with human perception and with VIF.

Next, we perform a simple clustering task where it is possible to evaluate the results manually. The *Heraldic Shields dataset* [6] (see Fig. 2) contains 12 images (of various sizes) which are to be clustered into 6 pairs. All possible pairwise distances are computed using the proposed distance measure. Hierarchical clustering is performed using the average linkage method. The clustering result shown in Fig. 2 demonstrates that our measure has discovered all 6 basic pairs of shields, and corresponds well with human intuition.

C. Clustering facial images

In this segment, we move towards more difficult clustering problems involving two larger benchmark datasets:

AT&T face [26]: This dataset contains 400 facial images of 40 individuals in 10 poses. These images (dimension: 112×92) are taken at different times with varying illumination, facial expressions and details.

Yale face [27]: This dataset has 165 grayscale facial images of 15 individuals. There are 11 images per subject, one per different condition: center light, with glasses, happy, left light, no glasses, normal, right light, sad, sleepy, surprised, and wink.

For each dataset, an $M \times M$ similarity matrix is computed using (9), where M is the number of elements in the dataset. This similarity matrix serves as the input to a standard spectral clustering algorithm [28]. The accuracy of the clustering results is measured using the Hungarian algorithm [29]. We compare our results with the compression-based state-of-the-art CK-1 distance measure [6] using the code provided by the authors. Due to the initialization process in spectral clustering, the accuracy varies slightly at each run. Figure 3 reports the mean clustering accuracies along with the standard deviations as computed over 10 runs for the two databases under consideration. The proposed measure outperforms CK-1 on the AT&T face dataset by 5.1% and its performance is 1.8% lower than CK-1 on the Yale dataset.

D. Texture retrieval

An image retrieval system, when provided with a query image, returns images from a large dataset that are perceptually similar to the query. We perform standard retrieval experiments on the following benchmark texture dataset.

Brodatz texture dataset [30]: This is a benchmark dataset that contains a variety of natural textures like grass and cloth. There are 111 different texture classes. Each original texture image is divided into 9 subimages to create the samples for that class.

For each query, the distances between the query and the remaining 998 images in the dataset are computed, and the first K nearest images are retrieved. The performance of a retrieval system is often measured in terms *Precision* and *Recall accuracy*. Precision is defined as the ratio of correctly retrieved images to the total number of images retrieved. Recall accuracy is defined as the ratio of the number of correctly retrieved images to the number of images available for the query class. Both precision and recall accuracy are expressed in terms of %. Our retrieval results are compared with those obtained using the CK-1 method in Fig. fig:retrv where our method clearly outperforms CK-1.

E. Texture classification

Supervised classification experiments are performed on a diverse collection of texture datasets drawn from the sources across various disciplines such as biology, medicine, forensics, etc.

UIUCTex [31]: This dataset features 25 texture classes with 40 samples each.

KTH Tips [32]: This dataset consists of textures of 10 different materials. The images vary in illumination, pose and scale.

Camouflage [6]: This dataset consists of 80 images of 9 varieties of modern US military camouflage. The images are created by photographing military t-shirts at random orientations.

Nematodes [6]: Nematodes are wormlike animals with great commercial and medical importance. Their species are often very difficult to distinguish from each other. This dataset contains 50 images of 5 different species of nematodes.

Tire tracks [6]: This is a collection of tire imprints left on a paper. It has 48 imprints of 3 different tires at varying directions.

Spiders [6]: This is a collection of images of Australasian ground spiders of the family Trochanteriidae. This family has high intra and inter-class variation.

VVT Wood [6]: This dataset contains 200 images of 40 types of wood defects (such as dry knot and small knot, etc.). The task is to label an image as either defective or sound.

TABLE I
CLASSIFICATION ACCURACY ON VARIOUS DATASETS OBTAINED USING THE PROPOSED DISTANCE MEASURE
AND THE STATE-OF-THE-ART COMPRESSION-BASED DISTANCE CK-1.

| Dataset | Classes | Proposed (%) | CK-1 [6] (%) |
|-------------|---------|--------------|--------------|
| Brodatz | 111 | 76.2 | 54.0 |
| UIUCTex | 25 | 51.6 | 51.0 |
| KTH Tips | 10 | 84.5 | 86.0 |
| Camouflage | 9 | 87.0 | 87.5 |
| Nematodes | 5 | 62.0 | 56.0 |
| Tire tracks | 3 | 79.2 | 79.2 |
| Spiders | 3 | 70.4 | 96.3 |
| VTT wood | 2 | 85.2 | 80.5 |

The classification results for the above datasets using the proposed method and the CK-1 are presented in Table I. We test both methods using a leave-one-out scheme in a 1-Nearest Neighbor framework. Our method demonstrates much better or comparable accuracy for all the datasets.

F. Discussion

Most compression-based methods use an off-the-shelf compressor (data, image or video compressor) and treat the compressor as a black-box. This makes it difficult to understand which part of the compression algorithm actually estimates the complexity of the data or measures the similarity. Consequently, the compression-based methods are difficult to improve upon, unless one wants to delve into the details of the compression algorithms. The proposed method takes a rather direct approach towards the approximation of complexity, and it is easier to understand and improve. Our method can be easily extended to measure the similarity between any type of signals including audio, video and other type of images such as medical images.

The proposed method requires learning a dictionary for each image. The dictionary learning process takes only a few seconds; for example, with the above-mentioned parameter values, a MATLAB implementation takes ~ 2 secs to learn a dictionary per image (including the patch extraction process) on a standard PC (intel quad @2.67GHz). This is as fast as any standard feature extraction process. However, our method is still slower compared to the compression-based CK1 measure. This can be explained by the fact that the areas of dictionary learning and sparse representation are still in the developing stage. In other words, unlike the standard compression algorithms, the existing algorithms for learning dictionaries

or sparse representations are not yet fully optimized for speed or memory.

We have used a greedy algorithm (OMP) to solve the sparse optimization problems in this work, primarily for speed and simplicity. Better results may be achieved using ℓ_1 regularized algorithms but at a higher computational cost. The proposed method is also not parameter-free, it requires a few parameters to be set by the user.

V. CONCLUSION

The main contribution of this work is the introduction of a sparse representation-based approach for computing a generic image similarity measure. The proposed measure has been shown to be successful in classifying, retrieving and clustering a variety of images as it performs consistently at par or better than the state-of-the-art. Nevertheless, the present work is not closed and we hope that this will stimulate interest in the areas of compression or Kolmogorov complexity-based similarity measurement using sparse representation.

A very recent work has also addressed the problem of similarity measurement using sparse representation of image features [33]. However, it addresses the problem from a different perspective and does not have any connection with the compression-based or Kolmogorov complexity-based approaches.

In this work, we have not focused on speeding up the classification, retrieval or the clustering processes since our objective has been to first demonstrate the usefulness and generality of the new distance measure. Future research will focus on using the measure more efficiently to classify and cluster larger datasets. This will require exploiting sophisticated machine learning techniques. Applications can also be extended to problems such as copy detection and data mining.

REFERENCES

- [1] B. Girod, "What's wrong with mean-squared-error?" *Digital Images and Human Vision*, 1993.
- [2] M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi, "The similarity metric," *IEEE Trans. Information Theory*, vol. 50, no. 12, pp. 3250 – 3264, Dec 2004.
- [3] R. Cilibrasi and P. Vitanyi, "Clustering by compression," *IEEE Trans. Information Theory*, vol. 51, no. 4, pp. 1523 – 1545, Apr 2005.
- [4] T. Watanabe, K. Sugawara, and H. Sugihara, "A new pattern representation scheme using data compression," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 579 –590, may 2002.
- [5] A. Pinho and P. Ferreira, "Image similarity using the normalized compression distance based on finite context models," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, sept. 2011, pp. 1993 –1996.
- [6] B. J. L. Campana and E. J. Keogh, "A compression-based distance measure for texture," *Statistical Analysis and Data Mining*, vol. 3, no. 6, 2010.

- [7] A. Macedonas, D. Besiris, G. Economou, and S. Fotopoulos, "Dictionary based color image retrieval," *J. Visual Comm. and Image Representation*, vol. 19, no. 7, pp. 464 – 470, 2008.
- [8] M. Li and Y. Zhu, "Image classification via lz78 based string kernel: A comparative study," in *Advances in Knowledge Discovery and Data Mining*, 2006, vol. 3918, pp. 704–712.
- [9] D. Cerra and M. Datcu, "A model conditioned data compression based similarity measure," in *Proc. DCC*, Mar 2008, p. 509.
- [10] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.
- [11] A. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems of information transmission*, vol. 1, no. 1, pp. 1–7, 1965.
- [12] R. J. Solomonoff, "A formal theory of inductive inference. part i," *Information and control*, vol. 7, no. 1, pp. 1–22, 1964.
- [13] G. J. Chaitin, "On the length of programs for computing finite binary sequences," *Journal of the ACM (JACM)*, vol. 13, no. 4, pp. 547–569, 1966.
- [14] M. Li and P. M. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, 2nd ed. Springer, 1997.
- [15] E. Keogh, S. Lonardi, and C. A. Ratanamahatana, "Towards parameter-free data mining," in *Proc. ACM SIGKDD*, 2004, pp. 206–215.
- [16] R. Cilibrasi, P. M. B. Vitányi, and R. de Wolf, "Algorithmic clustering of music based on string compression," *Computer Music Journal*, vol. 28, no. 4, pp. 49–67, 2003.
- [17] X. Chen, B. Francia, M. Li, B. McKinnon, and A. Seker, "Shared information and program plagiarism detection," *IEEE Trans. Information Theory*, vol. 50, no. 7, pp. 1545 – 1551, July 2004.
- [18] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.
- [19] Y. Pati, R. Rezaifar, and P. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar Signals, Systems and Computers*, 1993.
- [20] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Scientific Computing*, vol. 20, pp. 33–61, 1998.
- [21] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, pp. 4311–4322, 2006.
- [22] A. Tversky, "Features of similarity," *Psychological Review*, vol. 84(4), pp. 327–352, 1977.
- [23] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. of Computer Vision*, vol. 40, pp. 99–121, 2000.
- [24] B. Wohlberg, "Noise sensitivity of sparse signal representations: reconstruction error bounds for the inverse problem," *IEEE Trans. Signal Processing*, vol. 51, no. 12, pp. 3053 – 3060, 2003.
- [25] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Tran. Image Processing*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
- [26] [Online]. Available: <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>
- [27] [Online]. Available: <http://cvc.yale.edu/projects/yalefaces/yalefaces.html>
- [28] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. NIPS*. MIT Press, 2001, pp. 849–856.
- [29] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*. Dover Publications, 1998.

- [30] [Online]. Available: <http://www.ux.uis.no/~tranden/brodatz.html>
- [31] [Online]. Available: http://www-cvr.ai.uiuc.edu/ponce_grp/data/
- [32] [Online]. Available: <http://www.nada.kth.se/cvap/databases/kth-tips/download.html>
- [33] L.-W. Kang, C.-Y. Hsu, H.-W. Chen, C.-S. Lu, C.-Y. Lin, and S.-C. Pei, "Feature-based sparse representation for image similarity assessment," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1019–1030, 2011.



(a) original image

PSNR = ∞ , VIF = 1

Proposed = 0



(b) contrast change

PSNR = 24.53, VIF = 1.50

Proposed = 0.17



(c) luminance change

PSNR = 15.97, VIF = 0.95

Proposed = 0.20



(d) white noise

PSNR = 31.95, VIF = 0.96

Proposed = 0.33



(e) lossy jpeg

PSNR = 28.47, VIF = 0.92

Proposed = 0.38



(f) unrelated image

PSNR = 13.21, VIF = 0.14

Proposed = 0.54

Fig. 1. The proposed distance measure correlates well with human perception and with the well-known VIF method that measures perceptual signal fidelity.

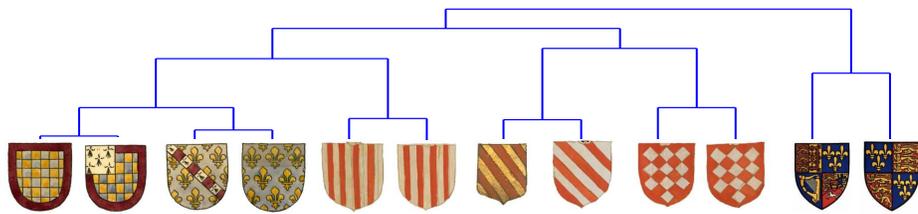


Fig. 2. Hierarchical clustering result on the Heraldic Shields dataset using the proposed sparse representation-based distance measure (although color images are shown here the result is obtained using grayscale images).

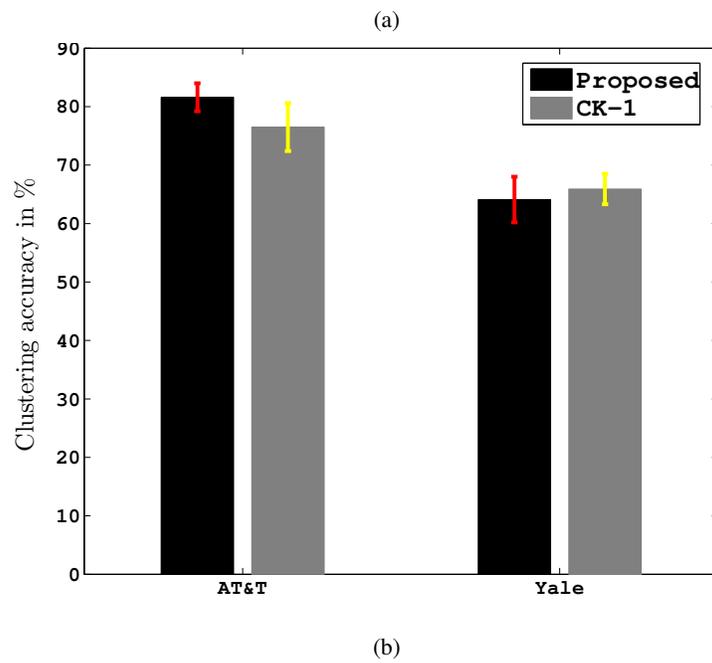


Fig. 3. (a) Sample images from the AT&T (first 3) and the Yale face (last 3) databases; (b) Clustering accuracy for the AT&T face (Proposed: $81.6 \pm 2.4\%$, CK-1: $76.5 \pm 4.1\%$) and the Yale face (Proposed: $64.1 \pm 3.9\%$, CK-1: $65.9 \pm 2.6\%$) databases.

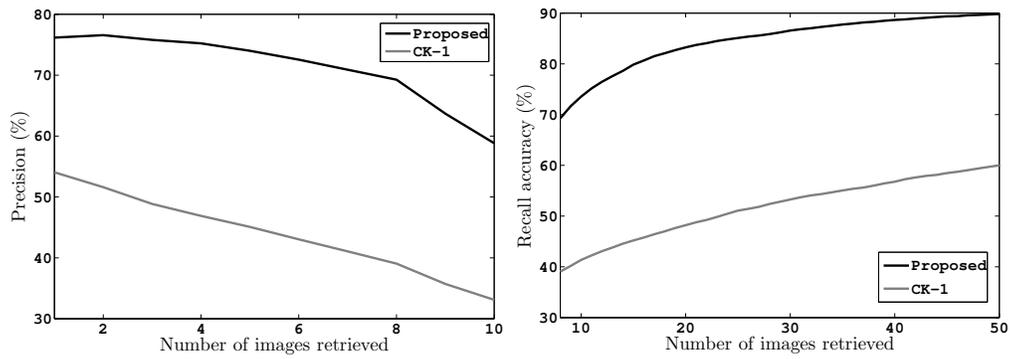


Fig. 4. Shown are the image retrieval results in terms of precision (left) and recall accuracy (right) obtained using the proposed method and the compression-based state-of-the-art CK-1 method on the Brodatz dataset.