IMAGE SIMILARITY MEASUREMENT FROM SPARSE RECONSTRUCTION ERRORS

Tanaya Guha, Rabab K Ward and Tyseer Aboulnasr

Electrical and Computer Engineering
The University of British Columbia, Vancouver, BC

ABSTRACT

This paper presents a new approach to measuring the similarity between two images using sparse reconstruction. Our approach alleviates the difficulty of selecting and extracting suitable features from images which usually requires domain-specific knowledge. The proposed measure, the *Sparse SNR* (SSNR), does not use any prior knowledge about the data type or the application. SSNR is generic in the sense that it is applicable, without modification, to a variety of problems involving different types of images. Given a pair of images, a set of basis vectors (dictionary) is learnt for each image such that each image can be represented as a linear combination of a small number of its dictionary elements. Each image is reconstructed by two dictionaries - the one trained on the image itself and the second - trained on the other image. We develop a novel similarity measure based on the resulting reconstruction errors. To the best of our knowledge, this is the first attempt to develop a sparse reconstruction-based similarity measure. Excellent classification, clustering and retrieval results are achieved on benchmark datasets involving facial images and textures.

Index Terms— image similarity, overcomplete dictionary, sparse reconstruction.

1. INTRODUCTION

A fundamental issue in image processing and understanding is that of measuring the similarity between a pair of images. Given the long history of similarity evaluation, numerous image similarity and distance functions exist in the literature - the simplest ones being the Mean Squared Error and the Euclidean distance. These measures compute the distance between images on a point-to-point basis and can not therefore interpret the visual appearance of images [1].

For problems like clustering, classification, retrieval, etc. where the results need to be compatible with the human notion of similarity, understanding the visual appearance of images is critical. This is often done by describing an image in terms of a set of *features* - a set of vectors with certain numerical attributes. The similarity between two images is then computed in terms of the similarity between their features. Although feature-based approaches are predominant, they suffer from a major limitation concerning the representational ability of the features. There are many domains (e.g. protein sequences) where it is not possible to find or even define satisfactory features. The features also need to be adapted to the application. Along with the features, the similarity measure also has to be changed or modified with the problem under consideration, e.g. the Mahalanobis distance is popular in comparing numerical vectors [2], while structured data like trees and graphs use the notion of edit distance [3], the earth mover's distance is useful for image retrieval [4], and the Euclidean distance works well for k-means clustering.

Recently, there has been a strong interest in developing *generic* similarity measures that are not specific to any application or data-type. One line of approach is to learn the similarity metric from the training data [2]. This method learns a Mahalanobis metric from a set of training samples. However, the method requires a good amount of training data in order to learn an effective similarity metric. Another approach relies on the information theory to quantify the complexity of one image in terms of the other [5, 6]. A practical algorithm based on this idea uses standard compression techniques to quantify how much information of one signal is contained in the other [6]. Although this method produces promising results for 1D discrete signals like text, its performance is not satisfactory for high dimensional signals such as images [7].

In this paper, we develop a generic similarity measure for images based on the theory of sparse signal representation. Several supervised classification algorithms based on the idea of sparse representation have been proposed in the past few years [8, 9, 10]. However, the problem of similarity measurement has not been addressed. To the best of our knowledge, this is the first attempt to develop a sparse reconstruction-based image similarity measure.

The basic idea in sparse analysis is to represent a signal by a linear combination of a small number of basis functions. It is well known that this is possible for many natural signals, such as audio and images, as long as they have sparse representation w.r.t. a properly chosen transform domain, e.g. music signals can be represented by a small number of sinusoids as they are sparse in the Fourier domain. Similarly, many natural images are sparse in wavelets domain.

In practice, however, signals are often a mixture of several structures. To achieve sparsity in such cases, we must com-

bine multiple bases. This results into an *overcomplete dictionary* - a set of basis vectors, where the number of bases is greater than the dimensionality of the input. Since the bases are tailored to represent signals similar to the training samples, an overcomplete dictionary requires fewer basis vectors to represent an input compared to complete dictionaries. This leads to higher sparsity in the transform domain. Overcomplete representations are also robust to additive noise, occlusion and translation of the input signal [11].

Given a pair of images, an overcomplete set of basis functions (not necessarily orthogonal) is *learnt* from each image such that each image has a sparse representation w.r.t. the bases. This approach mimics the *human visual system*; it has been shown that the basis functions learnt this way are qualitatively similar to the receptive fields of the simple cells in the mammalian primary visual cortex (V1) [12].

Once the dictionaries are learnt, they are used to quantify how well one image can be reconstructed using the information of the other. Each image is reconstructed by two dictionaries: the one trained on the image itself and the second one trained on the other image. We develop a novel similarity measure based on the resulting reconstruction errors. We use the signal-to-noise ratio (SNR) between the original and the reconstructed images to compute a similarity score which we name the *sparse SNR* (SSNR). To demonstrate the generality of SSNR, we perform experiments involving various applications. Our purely similarity-based results are comparable or better than the state-of-the-art indicating its high potential.

2. THE PROPOSED SIMILARITY MEASURE

Let us consider a pair of images, X and Y, both in $\mathbb{R}^N$. A natural way of measuring the similarity between X and Y is to quantify how well one image can be represented using the information in the other image. The more similar the images, the better is the representation of one image in terms of the other. Following this idea, we first build a dictionary for each image by extracting its dominant local structures. A similarity measure is then developed to measure how accurately one image can be approximated using the dictionary of the other. The steps of our proposed approach is described below.

2.1. Patch extraction

In order to learn an overcomplete dictionary for an image a large number of overlapping patches of dimension $\sqrt{n} \times \sqrt{n}$ are extracted *randomly* from each image and used as training samples for dictionary learning. Ideally, one patch centered at every image pixel is to be extracted; but in practice, extracting any large number of patches is sufficient for learning a good dictionary. Every image patch is converted to a vector of length n. Let the two sets of patches extracted from two images X and Y respectively be $\mathbf{P}_X$ and $\mathbf{P}_Y$, where $\mathbf{P}_X, \mathbf{P}_Y \in \mathbb{R}^{n \times m}$ and m is the total number of patches extracted from each image.

2.2. Dictionary learning

The next task is to learn an overcomplete dictionary for each image using the patches as input. For X, the goal is to learn a dictionary $\mathbf{\Phi}_X \in \mathbb{R}^{n \times k}$ having k ($k > n$) atoms, such that each patch (column vector) $\mathbf{p}_{\mathrm{x_i}} \in \mathbf{P}_X$ can be represented as a linear superposition of no more than τ ($\tau << k$) dictionary atoms. This optimization problem is therefore framed as

$$\min_{\mathbf{\Phi_X}, \mathbf{q}_\mathrm{x}} \sum_{i=1}^{m} \|\mathbf{p}_{\mathrm{x_i}} - \mathbf{\Phi}_X \mathbf{q}_{\mathrm{x_i}}\|_2^2 \text{ s.t.} \forall i \, \|\mathbf{q}_{\mathrm{x_i}}\|_0 \leq \tau \quad (1)$$

where $\mathbf{q}_{\mathrm{x_i}} \in \mathbb{R}^k$ represents the coefficients of the sparse representation of $\mathbf{p}_{\mathrm{x_i}}$ by $\mathbf{\Phi_X}$ and $\|.\|_0$ is the ℓ_0 seminorm that counts the number of non-zero elements in a vector. Similarly, a dictionary $\mathbf{\Phi}_Y$ is learnt for image Y. Due to the presence of the ℓ_0 term, (1) becomes a non-convex optimization problem, solving which accurately is NP hard. Instead, approximate solutions are found using greedy algorithms e.g. *orthogonal matching pursuit* (OMP) [13] or by ℓ_1 optimization [14]. We employ a fast dictionary learning algorithm called K-SVD [15] that solves (1). It performs two steps at each iteration: (i) sparse coding and (ii) dictionary update. In the sparse coding step, $\mathbf{\Phi}_X$ is kept fixed and $\mathbf{q}_\mathrm{x}$ is computed using OMP. During the second stage, the atoms of $\mathbf{\Phi}_X$ are updated sequentially, allowing the relevant coefficients in $\mathbf{q}_\mathrm{x}$ to change as well. For details of this algorithm please refer to the original K-SVD paper [15].

2.3. Similarity Measure

In this section, we develop a measure of similarity between images X and Y. Recall that, their corresponding overcomplete dictionaries are $\mathbf{\Phi}_X$, $\mathbf{\Phi}_Y$. A dictionary of an image is learnt by adapting the dictionary elements to the local structures of an image. Consequently, the atoms get tailored to represent the structures similar to those in the training samples. If X and Y are similar in appearance i.e. they have similar local structures, then $\mathbf{\Phi}_X$ and $\mathbf{\Phi}_Y$ will also be similar. In this case, $\mathbf{\Phi}_X$ will approximate the structures in Y with high accuracy i.e. the error obtained while representing Y using the dictionary of X will be small. The same will be true for the pair X and $\mathbf{\Phi}_Y$.

Consider two sets of random patches $\mathbf{U}_X = \{\mathbf{u}_{\mathrm{x_i}}\}_{i=1}^{l}$ and $\mathbf{U}_Y = \{\mathbf{u}_{\mathrm{y_i}}\}_{i=1}^{l}$ (where $\mathbf{u}_{\mathrm{x_i}}$ and $\mathbf{u}_{\mathrm{y_i}}$ are in $\mathbb{R}^n$) extracted from X and Y respectively. These patches are different from the patches used in learning. Let $\widehat{\mathbf{U}}_X$ and $\overline{\mathbf{U}}_X$ be the closest approximation (with a predefined sparsity constraint) of $\mathbf{U}_X$ by $\mathbf{\Phi}_X$ and $\mathbf{\Phi}_Y$ respectively.

$$\min_{\mathbf{v}_\mathrm{x}} \sum_{i=1}^{l} \|\hat{\mathbf{u}}_{\mathrm{x_i}} - \mathbf{\Phi}_X \hat{\mathbf{v}}_{\mathrm{x_i}}\|_2^2 \text{ s.t.} \forall i \, \|\hat{\mathbf{v}}_{\mathrm{x_i}}\|_0 \leq \tau \quad (2)$$

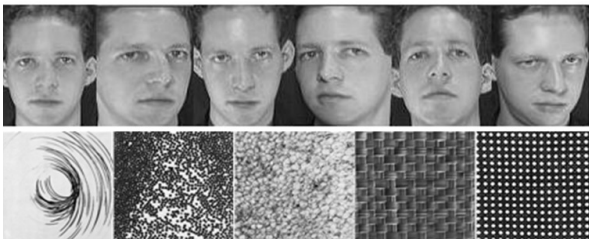

Fig. 1. Sample images from the AT& T face dataset (top row) and the Brodatz texture dataset (bottom row).

$$\min_{\mathbf{v}_x} \sum_{i=1}^{l} \|\overline{\mathbf{u}}_{x_i} - \mathbf{\Phi}_Y \overline{\mathbf{v}}_{x_i}\|_2^2 \text{ s.t. } \forall i \, \|\overline{\mathbf{v}}_{x_i}\|_0 \leq \tau \quad (3)$$

Similarly, $\widehat{\mathbf{U}}_Y$ and $\overline{\mathbf{U}}_Y$ are the closest approximations of $\mathbf{U}_Y$ by $\mathbf{\Phi}_Y$ and $\mathbf{\Phi}_X$. The proposed SSNR function $\mathbf{S}(X,Y) : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$ is defined as

$$\mathbf{S}(X,Y) = \frac{SNR\left(\mathbf{U}_X, \overline{\mathbf{U}}_X\right) + SNR\left(\mathbf{U}_Y, \overline{\mathbf{U}}_Y\right)}{SNR\left(\mathbf{U}_X, \widehat{\mathbf{U}}_X\right) + SNR\left(\mathbf{U}_Y, \widehat{\mathbf{U}}_Y\right)} \quad (4)$$

where, SNR is the signal-to-noise ratio. The numerator in (4) compares $\mathbf{U}_X$ and $\mathbf{U}_Y$ with their cross-approximations $\overline{\mathbf{U}}_X$ and $\overline{\mathbf{U}}_Y$. If X and Y are similar, both terms in the numerator will have large value. On the other hand, with significantly different X and Y such cross-approximations will produce low SNR values since $\mathbf{\Phi}_X$ and $\mathbf{\Phi}_Y$ will be quite dissimilar. The numerator is therefore a measure of similarity between two images. The more similar the images, the larger is the numerator. The denominator works as a normalizing factor. Note that,

$$\begin{aligned} SNR\left(\mathbf{U}_X, \overline{\mathbf{U}}_X\right) &\leq SNR\left(\mathbf{U}_X, \widehat{\mathbf{U}}_X\right) \\ SNR\left(\mathbf{U}_Y, \overline{\mathbf{U}}_Y\right) &\leq SNR\left(\mathbf{U}_Y, \widehat{\mathbf{U}}_Y\right) \end{aligned} \quad (5)$$

The above inequalities imply that $\mathbf{S}(X,Y) \leq 1$. The highest value of 1 is achieved when $X = Y$. The proposed SSNR has the following properties:
Non-negativity: $0 < \mathbf{S}(X,Y) \leq 1$. The similarity of an image to itself is one i.e. $\mathbf{S}(X,Y) = 1$ only when $X = Y$.
Symmetry: $\mathbf{S}(X,Y) = \mathbf{S}(Y,X)$. Symmetry is important because many clustering algorithms (e.g. spectral clustering) rely on this property.

3. PERFORMANCE EVALUATION

The generality of SSNR is demonstrated through four different applications: (i) facial image clustering, (ii) face recognition, (iii) texture classification and (iv) retrieval. For validation, two benchmark datasets (AT&T face and Brodatz texture datasets, see Fig. 1) are used. To learn a dictionary from an image, a set of $3,000$ random patches of size 8×8 are extracted from the image. Each trained dictionary (dimension 64×128) is learnt using $\tau = 8$ and 10 K-SVD iterations.

Table 1. Face recognition results on the AT&T dataset.

Approach	Recognition rate
Eigenface [18]	92.6%
ℓ_1 optimization [9]	93.3%
Mahalanobis [2]	97.4%
Proposed SSNR	**98.3%**

Table 2. Classification Results on the Brodatz texture dataset.

Approach	Classification accuracy
Affine invariant [19]	**87.4%**
Gabor filter [20]	85.1%
Proposed SSNR	86.3%

3.1. Clustering:

Clustering is the problem of automatically discovering the labels when unlabeled data is provided to the system. We perform clustering on the *AT&T face dataset*. This is a benchmark dataset that contains 400 grayscale images of 40 individuals in 10 poses. The images are collected at different times, with varying illumination, facial expressions and details. For clustering, a 400×400 similarity matrix is computed using SSNR. This matrix serves as the input to a standard spectral clustering algorithm [16]. The mean clustering accuracy for our method is **79.7%** with a standard deviation of 3.4%. This result is much superior to the correlation-based clustering that yields 68.5% mean accuracy and a standard deviation of 2.8%. The accuracy of the clustering results is measured using the Hungarian algorithm [17].

3.2. Face Recognition:

Face recognition experiment is performed on the AT&T face dataset. At each run, a training set is constructed by randomly selecting 7 images per class and the remaining 3 are used for testing. Classification is performed in a 3-Nearest Neighbor (3NN) framework. Table 1 presents the recognition results. Our result is compared with the state-of-the-art methods like ℓ_1-based classification approach [9] and the metric learning approach [2]. The result from the Eigenface [18] is used as baseline. Our results show improvement over all these methods.

3.3. Texture classification:

The *Brodatz texture dataset* used is the best-known texture dataset in literature. This is a highly diverse set of textures,

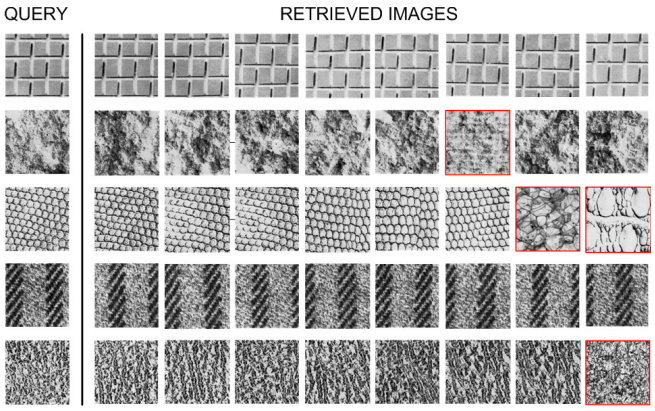

Fig. 2. Retrieval results for five query images from the Brodatz dataset. Eight nearest images are retrieved in each case. The images bordered in red, although perceptually similar to the query, belong to a different class from the query.

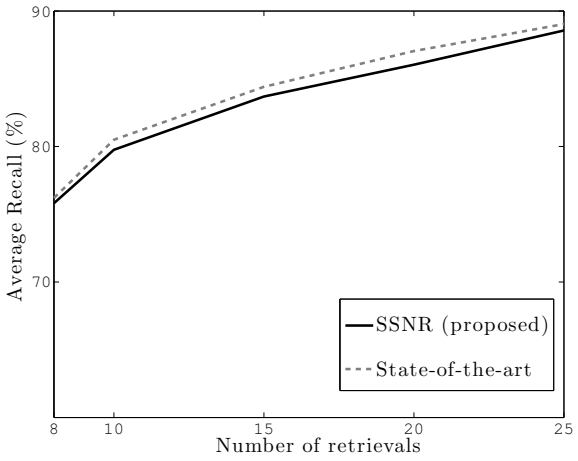

Fig. 3. Shown are the image retrieval results using SSNR and the state-of-the-art [19] on the Brodatz texture dataset. Our method closely follows the state-of-the art results: for number of retrievals = 8, the proposed SSNR has the average recall accuracy of 75.9%, which is comparable to the state-of-the-art accuracy of 76.2%. As the method in [19] is specifically designed for texture classification and retrieval, the performance of SSNR is quite encouraging.

some of which are perceptually quite close, while others are so inhomogeneous that it is very difficult, even for a human observer, to group their samples correctly. Following the methodology in [19], every image in each of the 111 texture classes is subdivided into 9 images of size 128×128. These 9 images are considered as samples of the same class. The training set is constructed by randomly selecting 3 images per class and the rest are used for testing. We use 1NN classification and 10 fold cross-validation. The results in Table 2 show that our classification accuracy is comparable with the current state-of-the-art. Interestingly, the methods that we compare with were specifically developed for texture classification (with texture-specific features) while our measure is generic and does not use any texture-specific features.

3.4. Retrieval:

A retrieval system, when provided with a query image, returns images from a large dataset that are perceptually similar to the query. For each query, SSNR is used to compute the similarity between the query and the remaining 998 images in the Brodatz dataset. The K nearest images with highest similarity are retrieved. Fig. 2 presents the retrieval results for 5 different query images where $K = 8$. For query 1 and 4, all the retrieved images belong to the class of the query yielding an retrieval accuracy of 100%. For query 2 and 4, 1 out of the 8 retrieved images does not belong to the class of the query i.e. the accuracy is 87.5%. Query 3 has 75% accuracy.

The performance of a retrieval system is often measured in terms of *average recall* which is defined as the ratio of the number of correct retrievals to the number of images available for the query class, expressed in terms of $\%$. For example, in the Brodatz dataset each class contains 9 images, when one is used as query, 8 are available for retrieval. If 10 nearest images are retrieved and 4 out of those 10 belong to the query class, the recall for that query is 50%. Average recall is computed by averaging over all queries. Our average recall result is compared with [19] in Fig. 3 where SSNR-based result closely follows the result of the state-of-the-art [19].

4. CONCLUSION

The main contribution of this work is developing a new image similarity measure (SSNR) based on the theory of sparse signal reconstruction. The advantage of this measure is that it does not use any prior knowledge about the application or the images being used. This alleviates the difficulty of selecting and extracting suitable features which often require domain-specific knowledge. SSNR is shown to produce state-of-the-art results for four applications involving images that are very different in nature (texture and faces). Our experimental results are purely based on similarity i.e. no sophisticated machine learning techniques have been used. Combining SSNR with powerful machine learning techniques will improve speed and accuracy further. SSNR can also be easily extended to work for color images and videos.

5. ACKNOWLEDGEMENT

This work was supported by NSERC Canada and by Qatar National Research Fund (QNRF No. NPRP 09-310-1-058).

6. REFERENCES

[1] B. Girod, "What's wrong with mean-squared-error?," *Digital Images and Human Vision*, 1993.

[2] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. NIPS*, 2006.

[3] M. Bernard, L. Boyer, A. Habrard, and M. Sebban, "Learning probabilistic models of tree edit distance," *Pattern Recognition*, vol. 41(8), pp. 2611–2629, 2008.

[4] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. of Computer Vision*, vol. 40, pp. 99–121, 2000.

[5] Ming Li, Xin Chen, Xin Li, Bin Ma, and P.M.B. Vitanyi, "The similarity metric," *IEEE Trans. Information Theory*, vol. 50, no. 12, pp. 3250 – 3264, Dec 2004.

[6] R. Cilibrasi and P.M.B. Vitanyi, "Clustering by compression," *IEEE Trans. Information Theory*, vol. 51, no. 4, pp. 1523 – 1545, Apr 2005.

[7] N. Tran, "The normalized compression distance and image distinguishability," in *Proc. SPIE*, 2007, vol. 6492.

[8] G. Peyré, "Sparse modeling of textures," *J. Math. Imaging Vis.*, vol. 34, no. 1, pp. 17–31, 2009.

[9] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. PAMI*, vol. 31, pp. 210–227, 2008.

[10] T. Guha and R.K. Ward, "Learning sparse representations for for human action recognition," *IEEE Trans PAMI*, vol. 34, pp. 1576 –1588, 2012.

[11] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[12] B.A. Olshausen and D. J. Field, "Natural image statistics and efficient coding," *Network: Computation in Neural Systems*, vol. 7, no. 2, pp. 333–339, 1996.

[13] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar Signals, Systems and Computers*, 1993.

[14] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, pp. 33–61, 1998.

[15] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. SP*, vol. 54, pp. 4311–4322, 2006.

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

[17] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Dover Publications, 1998.

[18] M. Turk and A. Pentland, "Eigen faces for recognition," *J. of Cognitive Neuroscience*, vol. 3, pp. 71–86, 1991.

[19] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Trans. PAMI*, vol. 27, no. 8, pp. 1265 –1278, aug. 2005.

[20] W.Y. Manjunath, B.S.and Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. PAMI*, vol. 18, no. 8, pp. 837 –842, aug 1996.