Signal Processing: *Image Communication* ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Image classification based on complex wavelet structural similarity ☆

Abdul Rehman *, Yang Gao, Jiheng Wang, Zhou Wang

Department of Electrical & Computer Engineering, University of Waterloo, Canada

ARTICLE INFO

Keywords:
Complex-wavelet structural similarity
Image classification
Handwritten digit recognition
Face recognition
Clustering
Support vector machine

ABSTRACT

Complex wavelet structural similarity (CW-SSIM) index has been recognized as a novel image similarity measure of broad potential applications due to its robustness to small geometric distortions such as translation, scaling and rotation of images. Nevertheless, how to make the best use of it in image classification problems has not been deeply investigated. In this paper, we introduce a series of novel image classification algorithms based on CW-SSIM and use handwritten digit recognition, and face recognition as examples for demonstration. Among the proposed approaches, the best compromise between accuracy and complexity is obtained by the CW-SSIM support vector machine based algorithms, which combines an unsupervised clustering method to divide the training images into clusters with representative images and a supervised learning method based on support vector machines to maximize the classification accuracy. Our experiments show that such a conceptually simple image classification method, which does not involve any registration, intensity normalization or sophisticated feature extraction processes, and does not rely on any modeling of the image patterns or distortion processes, achieves competitive performance with reduced computational cost.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Image classification is a common problem in a broad range of applications. The majority of existing image classification systems contains a "feature extraction" stage as a pre-classification step. These features are typically local or global structural descriptors of the image. The subsequent classification step then works in the feature space, where a large number of classifiers may be employed, ranging from simple *k*-nearest neighbor (*k*-NN) method [1] to more advanced approaches such as regularized discriminant analysis (RDA) [2], principle component analysis (PCA) mixture model[3], quadratic discriminant function (QDF) [4], and kernel-based support vector machine (SVM) [5] and kernel PCA methods [6]. The performance of these image classification systems is largely constrained by the extracted features, which need to be selected with great care, because "a classifier is only as good as its features". For example, since images or objects are often shifted, scaled and rotated, it is desirable to define (or design) the features so that they are invariant or robust to these changes [7]. There are also powerful machine learning algorithms, such as artificial neural networks [8] and convolutionary neural networks [9,10], that can be employed to automatically "discover" good features from a large number of training images, where feature discovery is left to a "black box" that may be obscure and difficult to understand in intuitive ways. A limitation of these feature-

☆ Partial preliminary results of this work were presented at *IEEE International Conference on Image Processing*, Brussels, Belgium, September 2011.

* Corresponding author.
E-mail addresses: abdul.rehman@uwaterloo.ca (A. Rehman), yang.gao@uwaterloo.ca (Y. Gao), jiheng.wang@uwaterloo.ca (J. Wang), zhouwang@ieee.org (Z. Wang).

0923-5965/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.image.2012.07.004

based approaches is that the features are tuned to specific classification problems and are weak in their generalization capability. As a result, the features may have to undergo a new phase of design, training or selection when images with different shapes and structures are to be classified.

A different type of image classification methods are based on template matching, where the similarities between a test image and a set of templates are evaluated and used to determine the class label without employing any specific structural features of images. These approaches are conceptually simple and often exhibit strong generalization ability. However, the effectiveness of these methods relies heavily on the *image similarity measure* being employed.

Recently, there has been significant progress in the design of image similarity measures [11]. In particular, the structural similarity (SSIM) index [12] has been found to be a much better measure than the widely used mean squared error (MSE) in predicting perceptual image quality, where the similarity between a distorted and a perfect-quality reference images is used as an indicator of the quality of the distorted image. The philosophy behind SSIM is to distinguish between structural and non-structural distortions and treat them unequally, which is presumably what the human visual system (HVS) would do.

Despite the superior performance of SSIM over MSE, both of them are very sensitive to geometric image distortions such as small scaling, rotation, and translation. In image classification tasks, however, resistance to these distortions is crucial because it is a common practice that images are not perfectly aligned to each other before a similarity measure is computed. In order to remove this "defect" from SSIM while maintaining its advantages, the complex wavelet SSIM (CW-SSIM) index was proposed [13], which is based on the correlations of phase patterns measured in the complex wavelet transform domain. The construction of CW-SSIM has some interesting connections with several computational models that account for a variety of biological vision behaviors. These models include: (1) the involvement of bandpass visual channels in image pattern recognition tasks [14]; (2) the representation of phase information in primary visual cortex using quadrature pairs of localized bandpass filters [15]; (3) the computation of complex-valued product in visual cortex [16]; (4) the computation of local energy (using sums of squared responses of quadrature-pair filters) by complex cells in visual cortex [17]; and (5) the divisive normalization of filter responses (using summed energy of neighboring filter responses) in both visual and auditory neurons [18,19]. CW-SSIM has been shown to be a useful measure in a series of applications, including image quality assessment [20], line-drawing comparison [20], segmentation comparison [20], range-based face recognition [21] and palmprint recognition [22]. However, its use in image classification problems has not been deeply exploited [23]. The previous CW-SSIM based methods [21,22] used the CW-SSIM NN case only. These methods are mainly based on the CW-SSIM values calculated between a query image and all the images in a database. The performance of these methods showed the suitability of CW-SSIM as the similarity measure for image recognition problems. However, the application of these methods is very limited due to their high computational complexity. The current paper explores a much wider ranger of CW-SSIM based image recognition methods and attempts to overcome the limitation of CW-SSIM NN. The method proposed in [23] for image classification uses CW-SSIM indirectly by employing it as a kernel function and exhibits lower recognition as compared to the proposed method.

In this study, we investigate CW-SSIM as a novel image classification tool in the context of handwritten digit and face recognition. The robustness of CW-SSIM against small geometric distortions allows us to avoid extracting any structural features that are insensitive to these distortions or employing any preprocessing methods such as deskewing, spatial shift, scaling and rotation. A series of CW-SSIM based classification methods are introduced, including CW-SSIM k-NN, CW-SSIM weighted k-NN, CW-SSIM k-means, and CW-SSIM SVM. Among them, CW-SSIM SVM achieves the best balance between classification accuracy and computational complexity, and is divided into two stages. In the first stage, an unsupervised clustering method is employed to divide the training images into clusters, each of which is associated with a representative image. In the second stage, a supervised learning method based on SVM is used to maximize the classification accuracy. The performance improvement of CW-SSIM SVM is achieved with reduced computational complexity.

2. Image similarity measures

The performance of template matching-based image classification systems critically depends on the accuracy and robustness of the image similarity/dissimilarity measure being employed, which quantifies the closeness or departure in the image space between a query image and any selected image in the training database.

2.1. Mean squared error

The mean squared error (MSE) is the simplest and most widely used image dissimilarity measure [24]. For two N-pixel grayscale images $\mathbf{x}$ and $\mathbf{y}$ with intensity values $\{x_i | i=1,\ldots,N\}$ and $\{y_i | i=1,\ldots,N\}$, respectively, the MSE is calculated as

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2. \quad (1)$$

The MSE is easy to compute and has a number of desirable properties in real world applications, but it also suffers from several fundamental problems [24].

Consider the handwritten digits in Fig. 1, where image (a) is used as a reference and compared with every other image. Regarded as collections of pixel intensity values and compared using MSE, the images are very different. However, regarded as shapes/structures, they appear rather similar to a human observer. In such a situation, if we persist on using MSE, then we need to perform

Please cite this article as: A. Rehman, et al., Image classification based on complex wavelet structural similarity, Signal Processing-Image Communication (2012), http://dx.doi.org/10.1016/j.image.2012.07.004

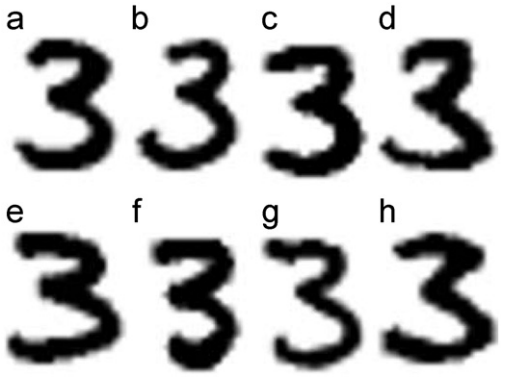

Fig. 1. Comparison of image similarity measures MSE and CW-SSIM. (a) Reference image; (b)–(h) test images with the same CW-SSIM but significantly different MSE values with respect to the reference: (a) MSE=0, CW-SSIM=1; (b) MSE=26, CW-SSIM=0.73; (c) MSE=17, CW-SSIM=0.72; (d) MSE=24, CW-SSIM=0.73; (e) MSE=15, CW-SSIM=0.73; (f) MSE=21, CW-SSIM=0.73; (g) MSE=33, CW-SSIM=0.72; and (h) MSE=14, CW-SSIM=0.72.

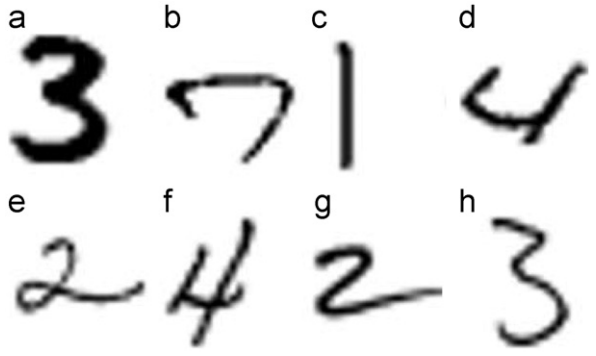

Fig. 2. Comparison of image similarity measures MSE and CW-SSIM. (a) Reference image; (b)–(h) test images with the same MSE but quite different CW-SSIM values with respect to the reference: (a) MSE=0, CW-SSIM=1; (b) MSE=57, CW-SSIM=0.29; (c) MSE=57, CW-SSIM=0.17; (d) MSE=57, CW-SSIM=0.33; (e) MSE=57, CW-SSIM=0.37; (f) MSE=57, CW-SSIM=0.33; (g) MSE=57, CW-SSIM=0.29; and (h) MSE=57, CW-SSIM=0.55.

various preprocessing steps and coordinate transformations to align the image patterns beforehand [25]. However, such alignment methods are often unreliable and any mis-registration of the images may lead to erroneous results. Another example is given in Fig. 2, where shapes or structures between the reference image (a) and each of the other images are substantially different, but their MSE values remain the same. Therefore, in order to operationalize the notion of shape/structure similarity, with ultimate goal of using it as a basis of a robust recognition system, we need to replace MSE with a similarity measure that is based on the similarity of shapes/structures between the images being compared.

2.2. Structural similarity indices

The SSIM index was originally proposed to predict human preference in evaluating image quality [12]. Assuming that the HVS is optimal in extracting structural information from the visual scene, a comparison of structural similarity should provide a good estimate of perceptual image similarities. The original SSIM algorithm works in the spatial domain. Given two image patches $\mathbf{x}=\{x_i|i=1,\ldots,M\}$ and $\mathbf{y}=\{y_i|i=1,\ldots,M\}$, the SSIM index is defined as

$$S(\mathbf{x},\mathbf{y})=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}, \tag{2}$$

where μ, σ are the sample mean, standard deviation or covariance, and C_1 and C_2 are two positive stabilizing constants, respectively [12]. The SSIM index is computed locally to compare image patches and then applied to an image using a sliding window approach followed by a spatial pooling stage [12]. The major drawback of the spatial domain SSIM index is its high-sensitivity to translation, scaling, and rotation of images [13,20], which are also non-structural distortions.

The CW-SSIM measure was proposed in [13,20], which was built upon local phase measurements in complex wavelet transform domain. The underlying assumptions behind CW-SSIM are that local phase pattern contains more structural information than local magnitude, and non-structural image distortions such as small translations lead to consistent phase shift within a group of neighboring wavelet coefficients. Therefore, CW-SSIM is designed to separate phase from magnitude distortion measurement and impose more penalty to inconsistent phase distortions.

Given two sets of complex wavelet coefficients $\mathbf{c}_x=\{c_{x,i}|i=1,\ldots,M\}$ and $\mathbf{c}_y=\{c_{y,i}|i=1,\ldots,M\}$ extracted at the same spatial location in the same wavelet subbands of the two images being compared, the local CW-SSIM index is defined as

$$\tilde{S}(\mathbf{c}_x,\mathbf{c}_y)=\frac{2\left|\sum_{i=1}^{M}c_{x,i}c_{y,i}^*\right|+K}{\sum_{i=1}^{M}|c_{x,i}|^2+\sum_{i=1}^{M}|c_{y,i}|^2+K}, \tag{3}$$

where c^* denotes the complex conjugate of c, and K is a small positive stabilizing constant. The value of the index ranges from 0 to 1, where 1 implies no structural distortion (but still could have small spatial shift). The global CW-SSIM index $\tilde{S}(I_x,I_y)$ between two images I_x and I_y is calculated as the average of local CW-SSIM values computed with a sliding window running across the whole wavelet subband and then averaged over all subbands. It was demonstrated that CW-SSIM is simultaneously insensitive to luminance change, contrast change, and small geometric distortions such as translation, scaling and rotation [13,20]. This makes CW-SSIM a preferred choice for image classification tasks because it is versatile and largely reduces the burden of preprocessing steps such as contrast and mean adjustment, pixel shifting, deskewing, zooming and scaling.

The performance of CW-SSIM is in clear contrast to that of MSE in the examples shown in Figs. 1 and 2. In Fig. 1, although there are notable variations in the spatial locations, orientations and thickness of the strokes in the test digit '3' images, they share similar structures, and consistently high CW-SSIM values are obtained, while there are significant differences in MSE values. In Fig. 2,

Please cite this article as: A. Rehman, et al., Image classification based on complex wavelet structural similarity, Signal Processing-Image Communication (2012), http://dx.doi.org/10.1016/j.image.2012.07.004

the test images represent different digits and have very different structures, but they share the same MSE value with respect to the reference image (a), making it impossible to select the right image (h) out of all test images. By contrast, CW-SSIM easily distinguishes image (h) among all test images because its CW-SSIM value is clearly the highest. These illustrative examples demonstrate the power of CW-SSIM, which does not require any pre-registration process but still provides consistently reasonable comparisons. This inspires us to use CW-SSIM as the image similarity measure in digit recognition and face recognition tasks.

3. CW-SSIM based image classification methods

Since CW-SSIM is a new similarity criterion introduced to the field, a series of image classification methods may be developed. In this section, we start from simple nearest neighbor algorithms to more sophisticated methods that lead to improved performance or reduced complexity. Here we present our algorithms with handwritten digit recognition and face recognition as our application in mind. However, the general approach is not restricted to this specific example, but should apply to many other applications as well.

3.1. CW-SSIM based nearest neighbor methods

Given a set of N training images $\{I_i|i=1,\ldots,N\}$ and their associated class labels $\{l_i|i=1,\ldots,N\}$ (in the case of digit recognition, there are 10 classes, each representing a digit between 0 and 9, i.e., $l_i \in [0,9]$), the most straightforward way of applying CW-SSIM for image classification is to find the image I_j in the training image set that is "closest" to a test query image I_q in CW-SSIM sense and use l_j to label the query image. This CW-SSIM based nearest neighbor (CW-SSIM NN) classifier can be expressed as

$$l(I_q)=l_j \quad \text{where } j=\arg\max_{i\in[1,N]} \tilde{S}(I_q,I_i). \tag{4}$$

Indeed, due to the desirable properties possessed by CW-SSIM, this conceptually simple algorithm can achieve very good performance, especially when the training set is large, as will be shown in Section 4.

The CW-SSIM NN classifier can be easily generalized to a CW-SSIM k-NN classifier. Given the k nearest neighbors of I_q (denoted by $\{I^{(i)}|i=1,\ldots,k\}$ and with class labels $\{l^{(i)}|i=1,\ldots,k\}$) in the training image set in terms of CW-SSIM, we use a majority vote to decide on the class label assigned to I_q

$$l(I_q)=\arg\max_{j\in[0,9]} \sum_{i=1}^{k} \delta(j,l^{(i)}), \tag{5}$$

where δ is a function such that $\delta(a,b)=1$ if $a=b$ and $\delta(a,b)=0$ otherwise.

The k-NN approach only considers the k-nearest neighbors of the query image in the full training image set. This can be interpreted as weighting the full sorted image set by a hard-weighting function which has value 1 for the first k images and 0 for the rest. It has been shown that soft-weighted k-NN can perform better than hard-weighted k-NN [26,27]. Therefore, we extend our CW-SSIM k-NN classification approach to a CW-SSIM weighted k-NN classifier, where the weight w_i is determined based on how close I_q is to the ith image in k neighbors

$$l(I_q)=\arg\max_{j\in[0,9]} \sum_{i=1}^{k} \tilde{S}(I_q,I^{(i)})\delta(j,l^{(i)}). \tag{6}$$

3.2. CW-SSIM based K-means method

A major problem with the nearest neighbor based methods described above is that they demand for CW-SSIM calculations of the query image with respect to all images in the training set. This could be computationally extremely expensive and thus prohibit its use in real-world applications. Classical methods for clustering such as k-means [28] have been used frequently in numerous vision applications [29]. Here we develop a CW-SSIM k-means clustering method to extract *typical* structures or *representatives* from the training image set and subsequently use these representatives to perform classification of the test query image with the help of nearest neighbor based methods.

k-means is an iterative algorithm that contains two steps in each iteration—updating the centroid for each cluster and updating cluster label for each sample image. Here we perform these two steps using CW-SSIM as the similarity criterion in replace of the typically used Euclidian distance. Given a set of R training images $\{I_i|i=1,\ldots,R\}$ that belong to a cluster, C, the centroid of the cluster is updated as

$$I_c, \quad \text{where } c=\arg\max_{i\in[1,R]} \sum_{j\in[1,R]} \tilde{S}(I_i,I_j). \tag{7}$$

Here the centroid I_c is not really the "mean" of all the images in the cluster, because CW-SSIM is not a valid distance metric in the image space and there is no simple definition of the notion of "mean" in terms of CW-SSIM. Rather, it is a representative image selected from all images in the cluster that on average is most similar to all other images in CW-SSIM sense. Given Z clusters with centroids $I_c^{(1)},I_c^{(2)},\ldots,I_c^{(Z)}$, the cluster label updating step is performed by reassigning the membership of each image I_i for $i=1,\ldots,N$ by

$$I_i \in C_j \quad \text{where } j=\arg\max_{j\in[1,Z]} \tilde{S}(I_i,I_c^{(j)}), \tag{8}$$

where C_j denotes the set of all images belonging to the jth cluster.

The above clustering algorithm group images in the training set without considering their class labels. As a result, a "bad" or outlier sample image may be clustered to a group of sample images that are similar in structure but have different class labels. To avoid such situations, for each training image I_t with class label l_t, we examine its k-nearest neighbors $\{I_t^{(i)}|i=1,\ldots,k\}$ with class labels $\{l_t^{(i)}|i=1,\ldots,k\}$ and compute the frequency of these neighbors that have the same class labels as I_t

$$\tilde{p}(I_t)=\frac{\sum_{i=1}^{k}\delta(l_t,l_t^{(i)})}{k}. \tag{9}$$

Please cite this article as: A. Rehman, et al., Image classification based on complex wavelet structural similarity, Signal Processing-Image Communication (2012), http://dx.doi.org/10.1016/j.image.2012.07.004

We then exclude the training images from the k-means clustering process with $\tilde{p}(I_t) < T_p$, where T_p is a preset threshold. Our experiments show that this training image pruning approach helps improve the classification results.

The k-means clustering process provides us with a set of cluster centroids or representative images. We can then apply the same weighted k-NN method for image classification as described in Eq. (6). The only difference is that the full training set used in Eq. (6) is replaced by the set of representative images. This leads to a much more efficient CW-SSIM weighted k-means image classification method.

3.3. CW-SSIM based support vector machine method

Motivated by the success of the SVM method [5] in a variety of pattern recognition tasks, we develop a CW-SSIM SVM image classification algorithm. The general structure of the algorithm is illustrated in Fig. 3, where the training phase consists of two main stages—an unsupervised clustering stage and a supervised SVM learning stage.

In the first stage, the training images are divided into clusters and one representative image (or template) is selected for each cluster. It is useful to be aware that there could be many different writing styles of the same digit, thus it makes sense to group the training images not only by their class labels, but also by their styles or structures. CW-SSIM is a preferred tool for this task than the existing similarity measures because images originated from the same digit and written with the same style are likely to be shifted, scaled, and/or rotated versions of each other. Our unsupervised clustering method works as follows. First, we calculate a matrix $\mathbf{C}$ of size $N \times N$, which contains the CW-SSIM values of every image with every other image in the training set. Each column of this matrix is a vector $\mathbf{s}_i = \{\tilde{S}(I_i, I_j) | j = 1, \ldots, N\}$ that contains the CW-SSIM values between the ith image and all other images in the training set. This vector may be considered as "features" of the ith training image (though not the descriptive features of image structures typically used in many other image classification methods). The clustering process starts by taking the whole training set as one cluster and defines the centroid of the cluster as

$$I_c^{(1)} \quad \text{where } c = \arg\max_{i \in [1,N]} \sum_{j \in [1,N]} \tilde{S}(I_i, I_j). \tag{10}$$

Now assume that we are at a stage where we have M clusters with centroids $I_c^{(1)}, I_c^{(2)}, \ldots, I_c^{(M)}$, respectively (the initial stage corresponds to $M=1$ case). We decide on whether to create a new cluster by checking whether

$$\min_{i \in [1,N]} \max_{j \in [1,M]} \tilde{S}(I_i, I_c^{(j)}) > T, \tag{11}$$

where T is a predefined threshold. If this is satisfied, then we can stop with the current number of clusters and use the corresponding centroids as representative images for the clusters. Otherwise, we define a new cluster centroid as

$$I_c^{(M+1)} = I_m \quad \text{where } m = \arg\min_{i \in [1,N]} \max_{j \in [1,M]} \tilde{S}(I_i, I_c^{(j)}), \tag{12}$$

and let $M = M+1$. After a new cluster is added, we reassign the membership of each image I_i for $i = 1, \ldots, N$ by

$$I_i \in C_j \quad \text{where } j = \arg\max_{j \in [1,M]} \tilde{S}(I_i, I_c^{(j)}), \tag{13}$$

where C_j is the collection of all images belonging to the jth cluster. The new centroid for each class $j \in [1,M]$ is then updated by

$$I_c^{(j)} = I_m, \quad \text{where } m = \arg\max_{I_i \in C_j} \sum_{I_k \in C_j} \tilde{S}(I_i, I_k). \tag{14}$$

This is followed by the next stage of judgement on whether a new cluster should be created, as in Eq. (11).

In the second stage of the training phase, we have the representative templates at hand. We can then describe any training image using a length-M vector of CW-SSIM values between the training image and all templates. Since every training image has a class label associated with it, this is a supervised learning problem. In particular, we develop a classifier by using support vector machines (SVM) with Gaussian kernels, which has been proven to be a powerful classifier of excellent generalization capability. Interested users can refer to [5] for details of the SVM learning algorithm.

The testing part of our CW-SSIM SVM classification algorithm is straightforward. For each test query image, we compute its CW-SSIM values with respect to all templates, resulting a length-M vector of CW-SSIM values. We then feed this vector to the SVM classifier, which produces a classification result.

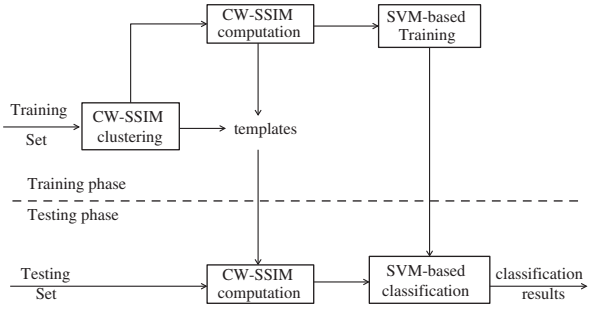

Fig. 3. Framework of the proposed CW-SSIM SVM method.

4. Experimental results

4.1. Handwritten digit recognition

Our experiments were performed on the MNIST database of handwritten digit images [30], which has been the most widely used benchmark in the literature. The database includes 60,000 training and 10,000 test samples. All images provided in the database have already been size-normalized and centered in a 28×28 box.

The performance of all the methods is given in Table 1 for different sizes of training set. Each experiment is performed five times with training data selected randomly for each experiment. The average values of the five experiments are presented. First, we compare the performance of MSE and CW-SSIM based nearest neighbor methods. The results for MSE NN and CW-SSIM NN

Please cite this article as: A. Rehman, et al., Image classification based on complex wavelet structural similarity, Signal Processing-Image Communication (2012), http://dx.doi.org/10.1016/j.image.2012.07.004

with different numbers of training images are shown in the second and third rows of Table 1. It appears that CW-SSIM alone, as a "raw" similarity measure (without any machine learning process involved), can achieve very good performance (less than 3% error rate) which is significantly better than the performance of MSE. As expected, the performance of CW-SSIM k-NN method can be improved when the values of $k>1$ are considered, as can be observed from the results in Table 1. Using CW-SSIM weighted k-NN (denoted as CW-SSIM (w) k-NN) helps us further improve the performance as the error rate reduces to 1.73% when the whole MNIST training data set is used for classification. Figs. 4 and 5 show the performance of CW-SSIM k-NN and CW-SSIM weighted k-NN, respectively, as a function of training set size for the values of $k=1,3,5,7$. It can be observed that the best performance is achieved for the value of $k=5$. Therefore, we use $k=5$ for all the experiments where k-NN is used as a classifier.

Second, we test the performance of, more practical, template based methods for different number of training images. For the results presented in Table 1, we learned 1150 representatives for each template based method. It can be observed that CW-SSIM pruned k-means (denoted as CW-SSIM (p) k-means) performs better than the method without pruning. The performance difference is higher for smaller sizes of training sets because pruning is expected to be more effective when the number of training images is lower in number. The value of the threshold, T_p, is set to be 0.5 which means that the training images that can not be correctly classified using the training set based on k-NN classifier are ignored. Test error rate of 4.56% suggests that the similarity measure helps us to achieve high accuracy even when a small fraction of training set is used for classification. Our CW-SSIM SVM algorithm outperforms aforementioned template based methods. An SVM is a binary classifier with discriminant function being the weighted combination of kernel functions over all training samples. For multi-class classification, binary SVMs are combined in either one-against-others or one-against-one (pairwise) scheme [31]. Note that in the clustering stage, the resulting number of clusters (and thus templates) varies with different choices of the threshold value T. The recognition error rate as a function of the number of templates is shown in Fig. 6. It can be observed that using a very small number of templates (38 out of 60,000 training images), the CW-SSIM SVM algorithm can achieve around 95% of accuracy. The error rate further decreases with the increasing number of templates, which collect more variations of representative structures. Some of the learned templates are shown in Fig. 7, where we can see that the templates are fairly different from each other even within each digit category, representing different writing styles.

The proposed CW-SSIM SVM method achieves lower error rate than the other two template based methods for

Table 1
Performance comparisons based on recognition error rate using MNIST database.

Training samples	2000 (%)	5000 (%)	10,000 (%)	20,000 (%)	30,000 (%)	60,000 (%)
MSE NN linear classifier [9]	12.57	10.41	9.56	8.23	7.62	6.92 8.4
2-layer NN MSE [9]						4.7
CW-SSIM NN	5.72	4.35	3.75	3.41	2.50	2.18
CW-SSIM k-NN	5.26	3.65	3.05	2.51	2.17	1.77
CW-SSIM (w) k-NN	5.08	3.57	2.95	2.39	2.12	1.73
CW-SSIM k-means	7.48	6.65	6.04	5.59	5.45	4.74
CW-SSIM (p) k-means	7.16	5.99	5.71	5.42	5.21	4.56
MSE SVM	10.22	7.61	6.31	5.15	4.60	4.16
CW-SSIM SVM	6.02	4.24	3.70	2.81	2.45	1.91
CW-SSIM (AP) SVM	5.00	3.93	3.46	2.71	2.40	1.89

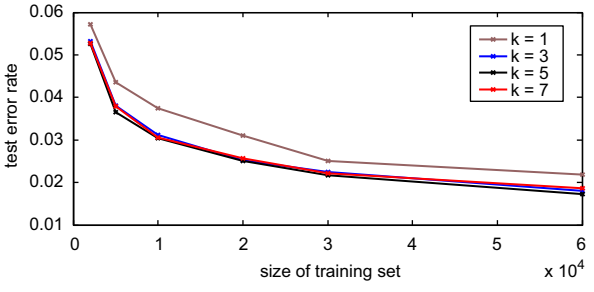

Fig. 4. Performance of CW-SSIM k-NN method as a function of training set size for different values of k.

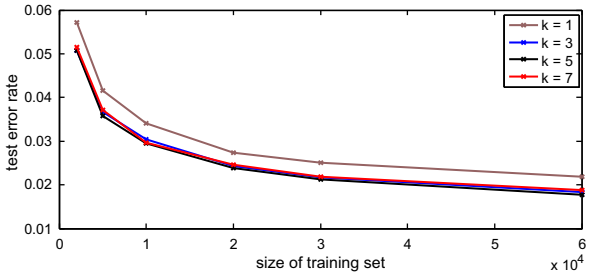

Fig. 5. Performance of CW-SSIM weighted k-NN method as a function of training set size for different values of k.

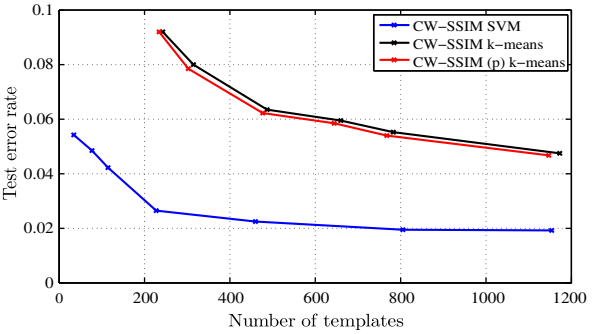

Fig. 6. Recognition error rate comparison of template-based proposed methods as a function of the number of templates.

Fig. 7. Sample templates learned from MNIST training set.

Table 2
Time saving by using CW-SSIM SVM as compared to CW-SSIM 1-NN.

Training samples	2000	5000	10,000	20,000	30,000	60,000
Time savings (%)	88.60	95.24	97.57	98.76	99.20	99.61

all the sizes of training set. The performance improves with the size of the training set. When all 60,000 training images are used, the error rate is reduced to less than 2%. It is important to mention that such improvement in recognition accuracy is obtained with largely reduced computational complexity because only a very small percentage of images (i.e., the selected templates) need to be compared as compared to the method that calculates CW-SSIM with all the images in the training set. As reported in Table 2, the time saving could be as high as 99.6%. Our non-optimal MATLAB implementation on a Intel Q9400 @ 2.66 GHz computer in single core mode takes about 2.5 s to classify a test image using 228 templates. It has the potential to achieve real-time performance with code optimization and hardware implementation.

The performance of the proposed method depends on the selection of templates. The use of better templates can lead to a better set of features for SVM that can result in higher classification accuracy. We use a recently proposed Affinity Propagation (AP) clustering method [32] for finding the templates, which has been shown to perform better than *k*-means based clustering methods. We refer to the classification algorithm with AP used for clustering as CW-SSIM (AP) SVM. Table 1 contains the performance of CW-SSIM (AP) SVM for different sizes of training set. Affinity propagation is an efficient algorithm, and it offers better templates, which means more complete and typical representation of all training images, for our later stage. The improvement of accuracy as compared to CW-SSIM SVM method decreases as the training set grows, because in that case, our former algorithm (described in Section 3.3) can achieve better results by increasing the number of clusters, which may finally cover almost all data points (used to be 30% more than affinity propagation). By using a relatively small number of templates, the SVM method with affinity propagation can achieve better classification accuracy.

Table 1 also compares the proposed approach with the following traditional approaches:

- MSE SVM
  - MSE SVM uses exactly the same method as the proposed method except for the use of MSE instead of CW-SSIM as the similarity measure. A significant difference between the performance of CW-SSIM SVM and MSE SVM can be observed.
- Linear classifier with deskewing as a pre-processing step [9].
- 2-Layer MSE based neural networks method [9].

Some of the misclassified digits are shown in Fig. 9. As can be observed that many of them are ambiguous and/or uncharacteristic, with obviously missing parts or strange strokes. Although there exist other recognition systems that achieved higher accuracy [30], they typically involve preprocessing stages (e.g., deskewing and denoising) and/or training and testing algorithms that are much more complicated in terms of both algorithm implementation and computational complexity.

4.2. Face recognition

In order to test the suitability of the proposed scheme as a general purpose image classification method, we employ it for face recognition. We used the proposed method to identify test images among 900 grayscale images as show in Fig. 8, available at [32], extracted from the Olivetti face database [33]. Olivetti database, or ORL database, contains a set of face images and was used in various face image applications. There are 10 different images from each of the 40 distinct subjects. For some subjects, the images were taken at different times, with varying lighting conditions, facial expressions (open or closed eyes, smiling or not smiling) and facial details (glasses or no glasses).

We compare the performance of four algorithms namely MSE SVM, CW-SSIM AP *k*-NN, CW-SSIM SVM and CW-SSIM (AP) SVM as shown in Table 3. Each experiment is performed five times with training data selected randomly from the training database, and the average values of five experiments are presented. Our CW-SSIM based image recognition methods are more efficient and accurate than other methods given in Table 3. It can be observed that CW-SSIM (AP) SVM uniformly achieves much lower error rate, and also as expected, save more than 60% computation time. It is also worth mentioning that with less than half of all data (400 training image out of 900, other 500 as testing set), the CW-SSIM (AP) SVM scheme can obtain a classification

Please cite this article as: A. Rehman, et al., Image classification based on complex wavelet structural similarity, Signal Processing-Image Communication (2012), http://dx.doi.org/10.1016/j.image.2012.07.004

Fig. 8. Samples of 900 images extracted from Olivetti database.

Fig. 9. Samples of misclassified test digits using proposed method. True label is given in the top right corner and the assigned label in given at the bottom of each image.

accuracy of more than 97.2% using fewer than 50 templates.

We then test the performance of our CW-SSIM (AP) SVM method with k-NN algorithm, with different values of k, and MSE SVM using the same templates clustered using affinity propagation. Table 3 shows that SVM significantly outperforms a series of k-NN methods, proving SVM is a much better tool to model training data behavior and produce competitive test results. In addition, our face recognition algorithm, implemented in MATLAB on a single-core, is sufficiently fast to achieve real-time performance.

Please cite this article as: A. Rehman, et al., Image classification based on complex wavelet structural similarity, Signal Processing-Image Communication (2012), http://dx.doi.org/10.1016/j.image.2012.07.004

Table 3
Performance comparisons based on recognition error rate using Olivetti face database.

Training images	900	800	700	600	500	400	300	200
Testing images	900	100	200	300	400	500	600	700
MSE SVM (%)	3.56	4.00	7.25	11.67	19.75	20.33	20.50	30.29
CW-SSIM AP k-NN (%)	3.11	5.00	5.50	7.33	12.25	20.75	23.83	29.14
CW-SSIM SVM (%)	1.34	2.25	3.33	4.47	6.25	7.75	10.76	15.23
CW-SSIM (AP) SVM (%)	0.00	0.00	0.00	1.33	2.00	2.80	6.25	10.14

5. Conclusion

We studied the problem of image classification using CW-SSIM as the image similarity criterion, which is connected with a number of computational models in biological vision and is robust to small geometric distortions of images. We use digit and face recognition as examples and propose a series of CW-SSIM based algorithms, which do not rely on any normalization, registration or image structure description-based feature extraction processes, and do not involve any statistical modeling of the image patterns or distortion processes, but achieve competitive performance in recognition accuracy. These properties make the proposed algorithms readily adaptable to a broad range of image classification problems.

Acknowledgment

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada and an Ontario Early Researcher Award, which are gratefully acknowledged.

References

[1] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13 (1) (1967) 21–27.
[2] J.H. Friedman, Regularized discriminant analysis, Journal of the American Statistical Association 84 (405) (1989) 165–175.
[3] H.C. Kim, D. Kim, S.Y. Bang, A numeral character recognition using the pca mixture model, Pattern Recognition Letters 23 (1–3) (2002) 103–111.
[4] F. Kimura, K. Takashina, S. Tsuruoka, Y. Miyake, Modified quadratic discriminant functions and the application to chinese character recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-9 (1) (1987) 149–153.
[5] C. Burges, A tutorial on support vector machines for pattern recognition, Knowledge Discovery and Data Mining 2 (1998) 1–47.
[6] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag New York, Inc., New York, NY, USA, 1995.
[7] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[8] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning Internal Representations by Error Propagation, MIT Press, Cambridge, MA, USA, 1986, pp. 318–362.
[9] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.
[10] P. Simard, D. Steinkraus, J. Platt, Best practices for convolutional neural networks applied to visual document analysis, in: Proceedings of the Seventh International Conference on Document Analysis and Recognition, 2003, pp. 958–963.
[11] Z. Wang, A.C. Bovik, Modern Image Quality Assessment, Morgan & Claypool Publishers, 2006.
[12] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612.
[13] Z. Wang, E.P. Simoncelli, Translation insensitive image similarity in complex wavelet domain, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, 2005.
[14] J.A. Solomon, D.G. Pelli, The visual filter mediating letter identification, Nature 369 (1994) 395–397.
[15] D.A. Pollen, S.F. Ronner, Phase relationships between adjacent simple cells in the cat, Science (212) (1981) 1409–1411.
[16] I. Ohzawa, G. DeAngelis, R. Freeman, Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors, Science (249) (1990) 1037–1041.
[17] E.H. Adelson, J.R. Bergen, Spatiotemporal energy models for the perception of motion, Journal of Optical Society of America 2 (2) (1985) 284–299.
[18] D.J. Heeger, Normalization of cell responses in cat striate cortex, Visual Neuroscience (9) (1992) 181–197.
[19] O. Schwartz, E.P. Simoncelli, Natural signal statistics and sensory gain control, Nature: Neuroscience 4 (8) (2001) 819–825.
[20] M.P. Sampat, Z. Wang, S. Gupta, A.C. Bovik, M.K. Markey, Complex wavelet structural similarity: a new image similarity index, IEEE Transactions on Image Processing 18 (11) (2009) 2385–2401.
[21] S. Gupta, M.P. Sampat, Z. Wang, M.K. Markey, A.C. Bovik, Facial range image matching using the complex wavelet structural similarity metric, in: Proceedings of the IEEE Workshop on Applications of Computer Vision, 2007.
[22] L. Zhang, Z. Guo, Z. Wang, D. Zhang, Palmprint verification using complex wavelet transform, in: Proceedings of the IEEE International Conference on Image Processing, 2007.
[23] G. Fan, Z. Wang, J. Wang, CW-SSIM kernal based random forest for image classification, Proceedings of the SPIE Visual Communications and Image Processing (2010).
[24] Z. Wang, A.C. Bovik, Mean squared error: love it or leave it? A new look at signal fidelity measures, IEEE Signal Processing Magazine 26 (1) (2009) 98–117.
[25] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (4) (2002) 509–522.
[26] J. Macleod, A. Luk, D. Titterington, A re-examination of the distance-weighted k-nearest neighbor classification rule, IEEE Transactions on Systems, Man and Cybernetics 17 (4) (1987) 689–696.
[27] J. Zavrel, An empirical re-examination of weighted voting for k-nn, in: Proceedings of the Seventh Belgian–Dutch Conference on Machine Learning, 1997, pp. 139–148.
[28] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, 1967, pp. 281–297.
[29] B. Moghaddam, A. Pentland, Probabilistic visual learning for object detection, in: International Conference on Computer Vision, 1995, pp. 786–793.
[30] Y. Lecun, C. Cortes, The MNIST database of handwritten digits, 2010. URL ⟨http://yann.lecun.com/exdb/mnist/⟩.
[31] U. Kressel, Pairwise Classification and Support Vector Machines, MIT Press, Cambridge, MA, USA, 1999, pp. 255–268.
[32] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (2007) 2007.
[33] F. Samaria, A. Harter, Parameterisation of a stochastic model for human face identification, in: Proceedings of the Second IEEE Workshop on Applications of Computer Vision, 1994, pp. 138–142.

Please cite this article as: A. Rehman, et al., Image classification based on complex wavelet structural similarity, Signal Processing-Image Communication (2012), http://dx.doi.org/10.1016/j.image.2012.07.004