IEEE International Conference on Image Processing (ICIP11), Brussels, Belgium, Sept. 2011

CW-SSIM BASED IMAGE CLASSIFICATION

Yang Gao , Abdul Rehman and Zhou Wang

Dept. of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada
Emails: yang.gao@uwaterloo.ca, abdul.rehman@uwaterloo.ca, zhouwang@ieee.org

ABSTRACT

Complex wavelet structural similarity (CW-SSIM) index has been proposed as a promising image similarity measure that is robust to small geometric distortions such as translation, scaling and rotation of images, but how to make the best use of it in image classification problems has not been deeply investigated. In this paper, we propose a novel "feature-extraction free" image classification algorithm based on CW-SSIM and use handwritten digit recognition as an example to demonstrate it. First, a CW-SSIM based unsupervised clustering method is used to divide the training images into clusters and to pick a representative image for each cluster. A supervised learning method based on support vector machines is then employed to maximize the classification accuracy based on CW-SSIM values between an input image and the representative images. 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 complexity.

Index Terms— complex-wavelet structural similarity, image classification, handwritten digit recognition, clustering, support vector machine

1. INTRODUCTION

Image classification is a common problem in a wide range of applications and involves both image processing and pattern recognition components. The majority of existing image classification systems contain a "feature extraction" stage as a pre-classification step. These features are structural descriptors of the image and need to be selected with great care, because "a classifier is only as good as its features". However, when image features are carefully designed and tuned to specific classification problems, they tend to become application-dependent and lose generalization capability. As a result, the features may have to undergo a new phase of design or training when images with different shapes and structures are to be classified. Certain machine learning techniques, such as artificial neural networks, could automatically "creates" features, but feature discovery is left to a "black box" that is obscure and hard to be understood in intuitive ways. On the other hand, template matching based classification methods, where the similarities between a test image and a set of templates are evaluated and used to determine the class label, are conceptually simple and require no sophisticated feature extraction processes. However, the effectiveness of such methods rely heavily on the *image similarity measure* being employed.

Recently, there has been significant progress in the design of image similarity measures. In particular, the structural similarity (SSIM) index [1] has been found to be a much better measure than the widely used mean squared error (MSE) in full-reference image quality assessment tasks, where the similarity between a distorted and a perfect-quality reference images is evaluated and 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 maintainng its advantages, the complex wavelet SSIM (CW-SSIM) index was proposed in [2], which has been shown to be a useful measure in a series of applications, including image quality assessment [3], line-drawing comparison [3], segmentation comparison [3], range-based face recognition [4] and palmprint recognition [5]. It has also been used for image classification tasks in a role of a kernel function [6].

In this study, we investigate CW-SSIM as a novel image classification tool in the context of handwritten digit recognition. Our method benefits from CW-SSIM as a powerful similarity measure that is robust against small geometric distortions. This allows us to avoid any preprocessing work such as deskewing, pixel shift, scaling, rotation and feature extraction. We show that CW-SSIM alone can achieve high performance but requires excessive computation because the CW-SSIM values between a test image and all images in the training set need to be calculated. To deliver a practical solution, the majority of our effort has been spent on learning

the most representative structures by employing CW-SSIM and on minimizing the classification error by using a support vector machines (SVM) based classifier [7] running on CW-SSIM values of the representative images. This results in improved performance with significantly reduced computational complexity.

2. METHOD

2.1. Structural Similarity Indices

The SSIM index was originally proposed to predict human preference in evaluating image quality [1]. 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 $\mathbf{x} = \{x_i | i = 1, ...M\}$ and $\mathbf{y} = \{y_i | i = 1, ...M\}$, the SSIM index is defined as:

$$S(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2\mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \qquad (1)$$

where μ, σ are the sample mean, standard deviation or covariance, and C_1 and C_2 are two positive stabilizing constants, respectively [1].

The major drawback of the spatial domain SSIM index is its high-sensitivity to translation, scaling, and rotation of images [2, 3], which are also non-structural distortions. To overcome this problem, the CW-SSIM measure was proposed in [2], 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 of 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. Specifically, given two sets of coefficients $\mathbf{c}_x = \{c_{x,i} | i = 1, ..., M\}$ and $\mathbf{c}_y = \{c_{y,i} | i = 1, ..., M\}$ 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:

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

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 [2, 3]. This makes CW-SSIM an ideal 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.

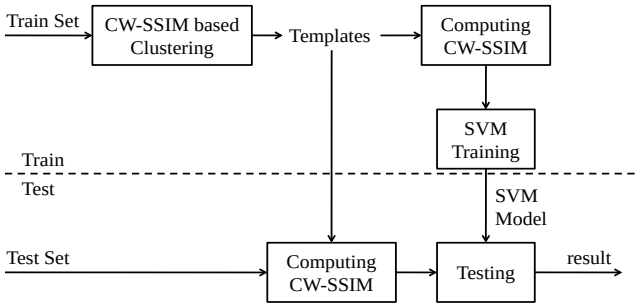

Fig. 1. Framework of the proposed method.

2.2. Proposed Method

Here we present our CW-SSIM based image classification method with handwritten digit recognition as our application in mind. However, the general approach we are presenting here is not restricted to this specific example, but should apply to many other applications as well. Given a set of N training images $\{I_i | i = 1, ..., N\}$ and their associated class labels $\{l_i | i = 1, ..., N\}$ (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 compute CW-SSIM between a test query image I_q and every training image I_i and then pick the one with the highest CW-SSIM value as the winner. This CW-SSIM alone algorithm can be expressed as

$$l_{recog}(I_q) = l_j, \text{ where } j = \arg\max_{i \in [1,N]} \tilde{S}(I_q, I_i). \qquad (3)$$

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 3. The problem with this method is that it demands for CW-SSIM calculations of the query image with the whole training set. This could be computationally extremely expensive and thus prohibit its use in real-world applications. There could also be other approaches based on decision rules different from Eq. (3) (e.g., K-nearest neighbors) that could achieve good recognition accuracy, but these approaches will suffer from the same complexity problem.

We propose a novel method with improved recognition performance but largely reduced complexity. The general structure is illustrated in Fig. 1. The training algorithm consists of two stages. 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 an ideal tool for this task 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 $\mathbf{s}_i = \{\tilde{S}(I_i, I_j)|j = 1, ...N\}$ that contains the CW-SSIM values between the i-th image and all other images in the training set. This vector may be considered as "features" of the i-th training image (though not descriptive features of image structures used in many other image classification methods). We start by taking the whole training set as one cluster and define the centroid of the cluster as

$$I_c^{(1)}\,,\ \text{where } c = \arg\max_{i\in[1,N]} \sum_{j\in[1,N]} \tilde{S}(I_i, I_j)\,. \qquad (4)$$

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

$$\min_{i\in[1,N]} \max_{j\in[1,M]} \tilde{S}(I_i, I_c^{(j)}) > T\,, \qquad (5)$$

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

$$I_c^{(M+1)} = I_k\,,\ \text{where } k = \arg\min_{i\in[1,N]} \max_{j\in[1,M]} \tilde{S}(I_i, I_c^{(j)})\,, \qquad (6)$$

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

$$I_i \in C_j\,,\ \text{where } j = \arg\max_{j\in[1,M]} \tilde{S}(I_i, I_c^{(j)})\,, \qquad (7)$$

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

$$I_c^{(j)} = I_m\,,\ \text{where } m = \arg\max_{I_i\in C_j} \sum_{I_k\in C_j} \tilde{S}(I_i, I_k)\,. \qquad (8)$$

This is followed by the next stage of judgement on whether a new cluster should be created, as in Eq. (5).

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 [7], which has been proven to be a powerful classifier of excellent generalization capability.

The testing part of our 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.

Table 1. Comparisons of Recognition Error Rate

Training Samples	MSE	CW-SSIM Alone	CW-SSIM +SVM	Time Saving
2000	12.57%	6.59%	6.02%	88.60%
5000	10.41%	4.68%	4.24%	95.24%
10000	9.56%	3.75%	3.70%	97.57%
20000	8.23%	2.99%	2.81%	98.76%
30000	7.62%	2.60%	2.45%	99.20%
60000	6.92%	2.38%	1.91%	99.61%

3. EXPERIMENTAL RESULTS

Our experiments were performed on the MNIST database of handwritten digit images [8], which has been the most widely used benchmark in the literature. The database includes 60,000 training and 10,000 test samples. All images have been size-normalized and centered in a 28×28 box.

First, we compare the performance of using MSE or CW-SSIM alone (as described in Section 2). The results with different numbers of training images are shown in the second and third columns in 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) and is significantly better than MSE.

Second, we test the effect of the number of templates on the performance of the proposed CW-SSIM + SVM algorithm. 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. 2. It can be observed that using a very small number of templates (38 out of 60,000 training images), the proposed 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. 3, where we can see that the templates are fairly different from each other even within each digit category, representing different writing styles.

Finally, we compare the proposed CW-SSIM + SVM algorithm with MSE or CW-SSIM methods alone. The results are shown in Table 1 for different numbers of training sam-

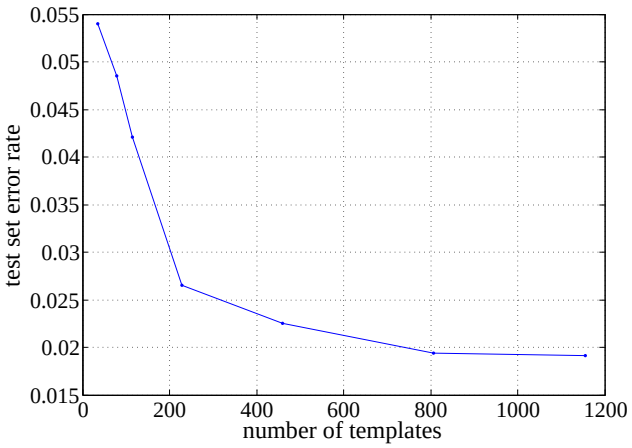

Fig. 2. Recognition error rate of proposed scheme as a function of the number of templates.

ples. It appears that in all cases, the proposed method achieves lower error rate than the other two methods. 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 reported in Table 1, the time saving could be as high as 99.6%. Our non-optimal MATLAB implementation on a Intel Q9400 @ 2.66GHz computer in single core mode takes about 2.5 seconds to classify a test image using 228 templates. It has the potential to achieve real-time performance with code optimization and hardware implementation. Although there exist other recognition systems that achieved higher accuracy [8], 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. CONCLUSION

We proposed a novel CW-SSIM based image classification method, which does not rely on any normalization, alignment or feature extraction processes, and does not involve any modeling of the patterns or distortion processes, but achieves competitive recognition accuracy with low computational complexity. These features make it a flexible approach that has good potentials to be applied to a broad range of image classification problems.

5. ACKNOWLEDGEMENT

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

Fig. 3. Sample templates learned from MNIST training set.

6. REFERENCES

[1] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[2] Zhou Wang and Eero P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Philadelphia, PA, Mar. 2005.

[3] Mehul P. Sampat, Zhou Wang, Shalini Gupta, Alan C. Bovik, and Mia K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Trans. Image Processing*, vol. 18, no. 11, pp. 2385–2401, Nov. 2009.

[4] Shalini Gupta, Mehul P. Sampat, Zhou Wang, Mia K. Markey, and Alan C. Bovik, "Facial range image matching using the complex wavlet structural similarity metric," *Proc. IEEE Workshop on Applications of Computer Vision*, Feb. 2007.

[5] L. Zhang, Z. Guo, Z. Wang, and D. Zhang, "Palmprint verification using complex wavelet transform," *Proc. IEEE Int. Conf. Image Proc.*, Sept. 2007.

[6] G. Fan, Zhou Wang, and Jiheng Wang, "CW-SSIM kernal based random forest for image classification," *Proc. SPIE Visual Comm. and Image Processing*, July 2010.

[7] Burges C., "A tutorial on support vector machines for pattern recognition," *Knowledge Discovery and Data Mining*, vol. 2, pp. 1–47, 1998.

[8] Y. LeCun and C. Cortes, *The MNIST Database*, http://yann.lecun.com/exdb/mnist.