21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Combining Multi-Scale Dissimilarities for Image Classification

Yan Li, Robert P.W. Duin, and Marco Loog
Pattern Recognition Laboratory, Delft University of Technology, The Netherlands
yan.li@tudelft.nl, r.duin@ieee.org, m.loog@tudelft.nl

Abstract

In image classification, multi-scale information is usually combined by concatenating features or selecting scales. Their main drawbacks are that concatenation increases the feature dimensionality by the number of scales and scale selection typically loses the information from other scales. We propose to solve this problem by the dissimilarity representation as it enables to combine various sources of information without increasing the dimensionality of the representation space. Various combining rules are introduced and tested with real-world applications. Our experiments show that combining with dissimilarities from all scales could indeed improve considerably upon the performance of the best single scale and adaptive combining can improve upon straightforward averaging.

1. Introduction

Incorporating multi-scale information can potentially improve the classification of images. Two common ways in the literature are concatenation of features at all scales [8, 4] and selection of scales [7]. The main drawback of concatenation is that it leads to very high dimensional feature vectors. Scale selection may provide the best scale for the problem, but it does not make full use of the image structure across the scales.

The dissimilarity representation [10] can be used to combine multi-scale information without increasing the dimensionality of the representation space. This representation characterizes an object by measuring its dissimilarities with a set of reference objects, called the representation set. It is especially useful when dealing with strings and graphs, or when the dissimilarity measure is not a distance or metric. The dissimilarity space is an easy-to-implement, yet very effective way for the dissimilarity representation [2]. Given a representation set of n images $R = \{\ell_1, \ell_2, \cdots, \ell_n\}$ and a dissimilarity measure d between images, the vector $D(\ell, R)$ for any image ℓ lies in the so-called dissimilarity space

$$D(\ell, R) = [\, d(\ell, \ell_1), d(\ell, \ell_2), \cdots, d(\ell, \ell_n)\,]. \quad (1)$$

This space can be treated as an Euclidean one and thus all statistical classifiers can be applied to it.

We propose to combine multi-scale information by (weighted) averaging or maximizing dissimilarities computed from all scales. Six such rules are introduced in Sect. 2, and tested with four datasets in Sect. 3. The experimental results are analysed in Sect. 4. The paper ends with a conclusion.

2. Combining Multi-Scale Dissimilarities

At each scale $\sigma_i, i = 1, \cdots, s$, the dissimilarities are computed when the image ℓ and all images in the representation set R are at scale i:

$$D^i(\ell, R) = [d(\ell^i, \ell_1^i), d(\ell^i, \ell_2^i), \cdots, d(\ell^i, \ell_n^i)]. \quad (2)$$

We consider the following six methods to combine those dissimilarities into a new one. The reasons for the particular choices are explained subsequently.

$$\begin{aligned}
D_a(\ell, R) &= \sum_{i=1}^{s} D^i(\ell, R);\\
D_b(\ell, R) &= \sum_{i=1}^{s} \frac{1}{\sigma_i} D^i(\ell, R);\\
D_c(\ell, R) &= \sum_{i=1}^{s} \frac{1}{\sigma_i^2} D^i(\ell, R);\\
D_d(\ell, R) &= \sum_{i=1}^{s} \alpha_i D^i(\ell, R);\\
D_e(\ell, R) &= \sum_{i=1}^{s} \alpha_i^4 D^i(\ell, R);\\
D_f(\ell, R) &= \max_{i=1,\ldots s} D^i(\ell, R).
\end{aligned} \quad (3)$$

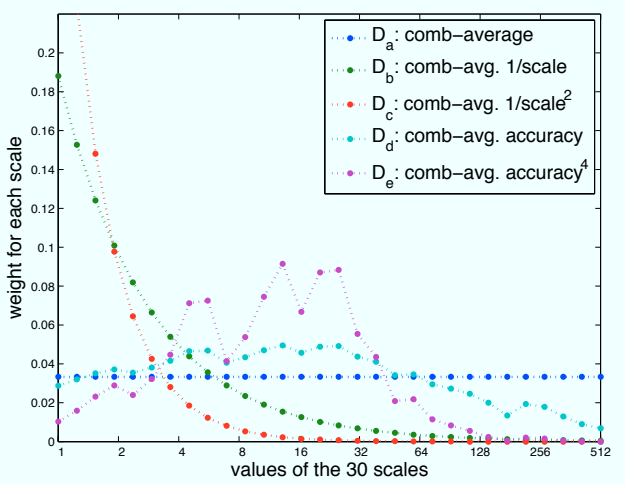

Figure 1. Weights for the 30 scales used in the colon tissue dataset. Note the difference between D_b and D_c, D_d and D_e.

The first five rules are weighted averages of dissimilarities. D_a is simply averaging dissimilarities from all scales. D_b assigns smaller weights for dissimilarities from larger scales. D_c is similar to D_b, but gives even less weights for large scales. D_d assigns weights according to the relative accuracy α_i. If the classification accuracy at scale σ_i is acc_i and the accuracy of random classification is rnd_{acc} (e.g., when the class priors are the same, rnd_{acc} is $1/n$ for a n-class problem), then α_i is defined as

$$\alpha_i = \max\left(acc_i - rnd_{acc},\, 0\right).$$

The higher the classification accuracy, the larger weight will be assigned to a scale. If the classification is random, then the weight becomes zero. D_e is similar to D_d, but takes the fourth power of α_i, which increases the gap between large and small weights[1]. If α_i is small, then α_i^4 will be very close to zero. D_f takes the maximum of the dissimilarities from all scales.

If we put dissimilarities from all scales into a vector,

$$V(\ell, \ell_j) = [d(\ell^1, \ell_j^1),\, d(\ell^2, \ell_j^2), \cdots,\, d(\ell^s, \ell_j^s)], \quad (4)$$

then D_a and D_f are related to L_1 and L_∞ norms of V:

$$D_a(\ell, \ell_j) = \frac{1}{s}\|V\|_1,\; D_f(\ell, \ell_j) = \|V\|_\infty.$$

While D_b and D_d have many weights of small values, their counterparts, D_c and D_e respectively, typically make those small weights close to zero. A justification of this operation may come from [11], which shows that combination of *some* good individuals can perform better than the combination of *all* individuals.

As an example, the weights for the colon tissue dataset (described in the next section) is shown in Figure 1. In comparison with D_b, D_c gives much higher weights to small scales, and about a half of the scales are assigned close-to-zero weights. While the weights for D_d are relatively flat across scales, those for D_e are much more different. Weights for scales larger than 128 become almost zero.

It should be mentioned that combining dissimilarities is not a new idea, see for example [9] for general investigations. The focus of this paper is to use dissimilarities to combine multi-scale information. In particular, combining rules D_b and D_c explicitly use the scale information for the combining, and D_d and D_e have not been studied for dissimilarities before.

3. Experiment Setup

We tested different combination rules on the following four datasets.

The first is the chicken pieces silhouettes dataset (*chicken*) [1]. It has 446 pieces from 5 classes. Pieces were represented as strings, and weighted edit distance were computed as dissimilarities. Eleven resolutions were used for the string representation, resulting in 11 dissimilarities of different scales.

The second is the MNIST digit (*MNIST-dig*) recognition [5], which is a ten-class problem classifying digits 0 to 9. Each digit had 300 examples sampled from the training digits. The Euclidean distance between pixel values was used to measure dissimilarities. The images were smoothed with Gaussian kernels of 10 different scales (standard deviations). Including the original images, a total of 11 dissimilarities were computed.

The third is the Brodatz textures (*Brodatz*), which have 111 images. Following the general test protocol, each image was partitioned into 9 subimages of size 215×215, resulting 999 images for classification. The Leung-Malik filter set [6] at each scale was applied to the images, and L_1 distance between the histograms of the response images were computed as dissimilarities. Twenty-one scales sampled exponentially between 1 and 32 were used.

The fourth comes from a colon tissue dataset (*colon*) acquired by Marius Nap of the Atrium Medical Center in Heerlen, The Netherlands. The objects are microscopy image patches of size 1000×1000, belonging to four classes: normal, inflamed, adenomatous, and cancer. The Laplacian Δ of different scales were

[1] The fourth power was used to make D_d and D_e sufficiently different. Other choices are possible and at times may be better.

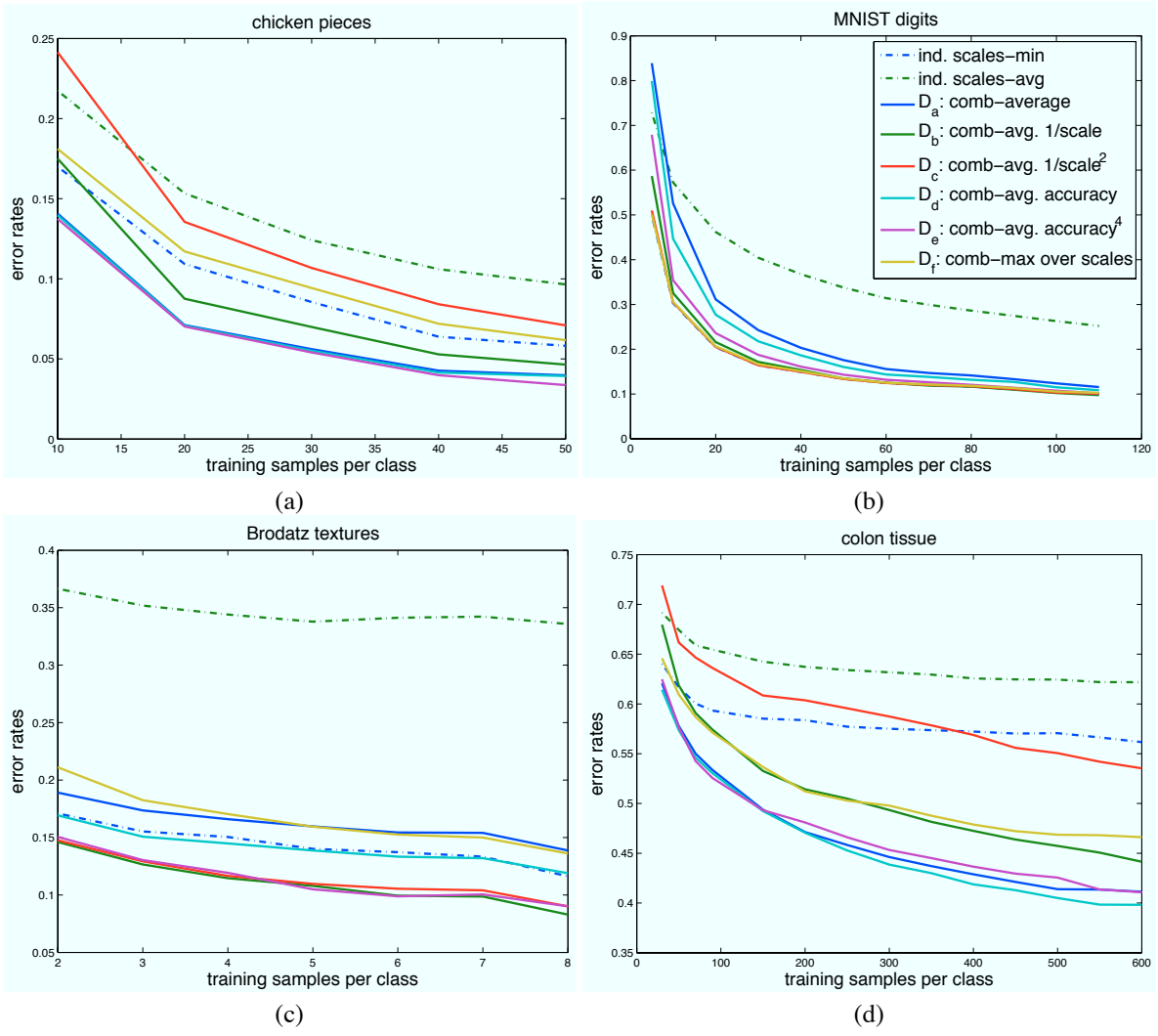

Figure 2. Learning curves for six methods to combine multi-scale dissimilarities (cf. Eq. (3)), the best individual scale, and the average performance of all scales. The same legends are used as in (b). In (b), the learning curve of 'ind. scales-min' overlaps with those of D_c and D_f.

applied to each image patch, and the Euclidean distance between the histograms of the response images was used as the dissimilarity measure.

The results of the combining rules, the average performance of all scales, and the best individual scale which has the minimum error rate on average, are shown in Figure 2. For each dataset, the same classifier was used for individual scales and for the combined dissimilarities. For *Brodatz* and *colon*, linear SVM with unit penalty parameter was used; for *chicken* and *MNIST-dig*, linear discriminant analysis was used, together with a principle component analysis for dimensionality reduction. Classifiers were chosen only to illustrate the combining rules; they were not optimized upon classification performances. The weights α_i for D_d and D_e were computed based on the cross-validation classification with all the data. For one dataset at each scale, all the dissimilarities were normalized to make their maximum equal to one. The experiments were repeated for 10 times, and the average error rates are reported here.

4. Analysis

The learning curves of individual scales (not included here) showed that for different datasets, the performances across scales differ significantly. The scale ranges in which good performance were achieved

also varied: small scales for *MNIST-dig* and *Brodatz*, medium scales for *colon*, and both medium and large scales for *chicken*. This variation influenced the behaviour of combining rules, as we will see later on.

From Figure 2, the first thing to notice is that combining multiple scales can improve upon the average performance of all scales. A proper combination could even improve upon the best individual scale significantly. Especially for *colon*, all combine rules except D_c outperformed the best individual for a large margin with more than 200 training samples per class. For *chicken* and *Brodatz*, the best combining methods also outperformed the best individual a lot. This justifies the combining of multiple scales, which can potentially improve performance upon individual scales.

With respect to different combining rules, the simple average rule D_a performed already fairly well. This is in agreement with the results in [9, 3]. For *chicken* and *colon*, D_a was among the best. The worse performance of D_a for *MNIST-dig* and *Brodatz* may come from the averaging of a large percentage of scales which performed much worse than the best scale. How to combine the scales of low performance is that task of adaptive combining methods.

The combining rules D_b and D_c assign smaller weights to larger scales. They were effective when larger scales lead to worse results, as were the case for *Brodatz* and *MNIST-dig*. For both datasets, D_b and D_c achieved performance among the best of all combining rules. It should be noted that those larger scales, though much worse than the best scale, could still provide valuable information. See for example *Brodatz*, D_b outperformed the best individual with a large margin. When the best performance are achieved at medium or large scales, however, D_b and especially D_c tend to work bad. This explains why D_c was the worst combining rule for *chicken* and *colon*. D_b was better than D_c since it did not penalize that much for large scales, but was not among the best combining rules.

The combining rules D_d and D_e assign weights to scales according to their classification errors. For all four datasets, D_e was always among the best combining rules, and outperformed upon, or was comparable to, the best individual scale. D_d also performed very well on *chicken* and *colon*, but was worse than D_e on *MNIST-dig* and *Brodatz*. This indicates that 'stretching' the weights by assigning larger weights to good scales and close-to-zero weights to very bad scales is useful.

Combining by taking the maximum dissimilarity across scale, D_f, was among the best only for the MNIST digits dataset. Though it can be interpreted as a different norm as D_a, their performances differed a lot. A deeper understanding of D_f is still to be explored.

5. Conclusions

We have tested six rules to combine dissimilarities from different scales. Our experiments show that (1) Combining multi-scale information with the dissimilarity representation could improve upon the best individual scale. (2) Combining by averaging the dissimilarities across scales can already provide good classification results. (3) Weighting dissimilarities according to the classification accuracy at each scale is a good candidate for choosing weights. (4) When the classification results at large scales are worse than those at small scales, dissimilarities may be combined by assigning weights inversely proportional to scale values. The large scales still provide valuable information, which can potentially improve the best individual scale.

6. Acknowledgement

The authors thank Dr. Marius Nap of the Atrium Medical Center in Heerlen, the Netherlands, for providing the colon tissue dataset.

References

[1] H. Bunke and U. Bühler. Applications of approximate string matching to 2d shape recognition. *Pattern recognition*, 26(12):1797–1812, 1993.

[2] R. Duin and E. Pekalska. The dissimilarity space: bridging structural and statistical pattern recognition. *Pattern Recognition Letters*, 33(7):826 – 832, 2012.

[3] A. Ibba, R. Duin, and W. Lee. A study on combining sets of differently measured dissimilarities. *ICPR*, 2010.

[4] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

[6] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, 2001.

[7] T. Lindeberg. Feature detection with automatic scale selection. *Int. J. Comput. Vis.*, 30(2):79–116, 1998.

[8] J. Mao and A. Jain. Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern recognition*, 25(2):173–188, 1992.

[9] E. Pekalska and R. Duin. On combining dissimilarity representations. *Multiple Classifier Systems*, 2001.

[10] E. Pkkalska and R. Duin. *The dissimilarity representation for pattern recognition*. World Scientific, 2005.

[11] Z. Zhou, J. Wu, and W. Tang. Ensembling neural networks: many could be better than all. *Artificial intelligence*, 137(1):239–263, 2002.