Kernel Methods for Extracting Local Image Semantics

Ben Bradshaw
Bernhard Schölkopf
John C. Platt

October 2001

Technical Report
MSR-TR-2001-99

This paper describes an investigation into using kernel methods for extracting semantic information from images. The specific problem addressed is the *local* extraction of 'man-made' vs 'natural' information. Kernel linear discriminant and support vector methods are compared to the standard linear discriminant using a multi-level hierarchy. The two kernel methods are found to perform similarly and significantly better than the linear method. An advantage of the kernel linear discriminant over the SVM method is that accurate class-conditional density estimates can be determined at each level allowing posterior estimates of class membership to be evaluated. These probabilistic outputs give a principled framework for combining results from a number of semantic labels.

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
http://www.research.microsoft.com

1 Introduction

This paper investigates architectures for generating local semantic labels within an image. The motivation for this problem is in content-based image retrieval. Once the spatial layout of a number of semantic features can be extracted, combinations of these can be used to formulate complex image retrieval queries based on the *spatial semantics* of the underlying images. Semantic query formulation leads to the representation of a much richer class of concepts than those of current retrieval systems, whose queries are based on the local outputs of low-level image features — for instance colour histograms or texture [1]. Finally note that, since the proposed method generates posterior probabilities, diverse semantic outputs can be combined in a principled manner.

There are a number of papers that address the issue of determining the semantic content of images, all of which do so at a global scale (i.e. they result in one output per image). The papers most similar to the work presented here are those of Torralba *et al.* [13] and Vailaya *et al.* [14]. The former paper describes an algorithm that attempts to determine a set of real-valued 'semantic axes' in a particular feature space. They recognise the importance of being able to assign real-values to each image in relation to each semantic label, rather than the more common binary classification approach, but do not extend these real-values to a probabilistic representation. The latter paper, by Vailaya *et al.* describes a system that performs a hierarchical categorisation of images using a Bayesian framework which results in probabilistic labels for the images.

All of the systems referenced above output only one binary or real value per image. The significant contribution of this paper is to illustrate that semantic labelling can be localised to a *small area* in an image (rather than the entire image). To achieve this various architectures and machine learning methods, described in Section 2, are tested to evaluate the optimum approach. Section 3 highlights the results from these methods and illustrate that the proposed algorithms can indeed be used to generate accurate localised semantic labels.

It should be stressed that this class of problem (i.e. classification into broad semantic categories) is very different to specific object-detection algorithms such as the face detectors proposed by Rowley [9] or Papageorgiou *et al.* [5]. There are much larger intra-class variabilities present when considering these broad semantic categories; to illustrate this, consider the problem of trying to locate cars in an image and compare it to the problem of trying to locate man-made objects in an image. It is clearly much easier to define a model of the former class than the latter class.

2 Proposed architectures

2.1 Sampling procedure

The proposed sampling procedure extracts texture and colour data from different sized blocks from the image each of which is centred at the current sampling point. Figure 1 illustrates this procedure. In this paper, samples are extracted using a 'grid' with a 16×16 pixel spacing. Blocks 128×128, 64×64, 32×32 and 16×16 pixels in size are used, these being denoted as levels 1, 2, 3 and 4 respectively. Note that all images

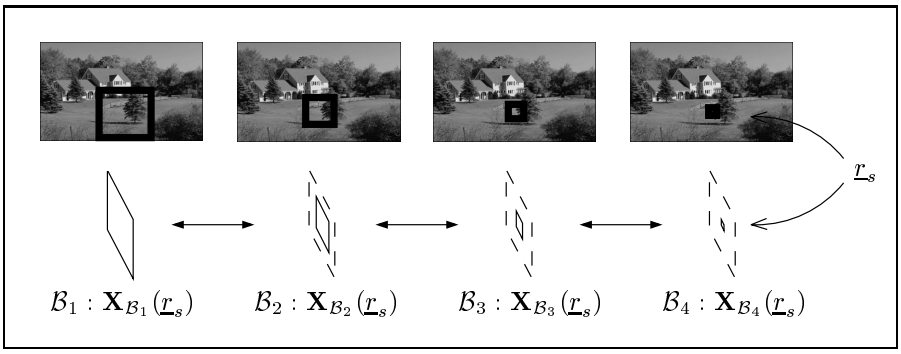

Figure 1: **Blocks contributing to a sample are all centred at the same pixel position.** Given sample position $\underline{r}_s$ a number of feature vectors are extracted. Each feature vector at level l, corresponding to block $\mathcal{B}_l$ is subsequently denoted as $\mathbf{X}_{\mathcal{B}_l}(\underline{r}_s)$.

used are either 256×384 pixels or 384×256 pixels in size, and have been extracted from the Corel Gallery 1,000,000 collection.

The feature vectors extracted from each block have 26 dimensions, 24 texture features and 2 colour features, all derived from an Ohta transformation of the original image [4]. To reduce the dimensionality of the feature space, textural features are only extracted from the luminance component, denoted I_{o1}, whilst colour features are extracted from the chrominance components, denoted I_{o2} and I_{o3}. Chrominance information at sample point $\underline{r}_s$ (in this paper $\underline{r}$ is used to denote pixel position) is obtained from block $\mathcal{B}_l$, (see Figure 1) as follows:

$$\mathcal{C}_1(\mathcal{B}_l) = \int_{\mathcal{B}_l} I_{o2}(\underline{r})\, \mathrm{d}\underline{r}, \qquad \mathcal{C}_2(\mathcal{B}_l) = \int_{\mathcal{B}_l} I_{o3}(\underline{r})\, \mathrm{d}\underline{r} \tag{1}$$

The texture features are extracted using the complex wavelet transform (CWT) which was developed by Kingsbury [2] and is an efficient way of implementing a set of critically sampled Gabor-like wavelets. Gabor wavelets/filters have been used by a number of authors investigating both semantic content and classification problems [10], [13]. The advantage of using the CWT rather than a Gabor wavelet method is the significantly reduced computational load; instead of requiring 2 dimensional convolutional operations the CWT combines results obtained from computationally efficient 1 dimensional convolutional operations.

The CWT wavelet function at scale s, and orientation θ is denoted as ϕ_s^θ. The orientation can take one of six values $\Theta = \{15°, 45°, 75°, -75°, -45°, -15°\}$. The θ in the following text refers to an index into this vector i.e. $\theta \in \mathcal{I} : \{1 \ldots 6\}$. The energy response across block $\mathcal{B}_l$ to the wavelet function at scale s, and orientation θ, when applied to the Ohta luminance image is defined as:[1]

$$\mathcal{T}_s^\theta(\mathcal{B}_l) = \int_{\mathcal{B}_l} (I_{o1}(\underline{r}) * \phi_s^\theta)^2\, \mathrm{d}\underline{r} \tag{2}$$

[1]The $*$ symbol denotes the convolution operator

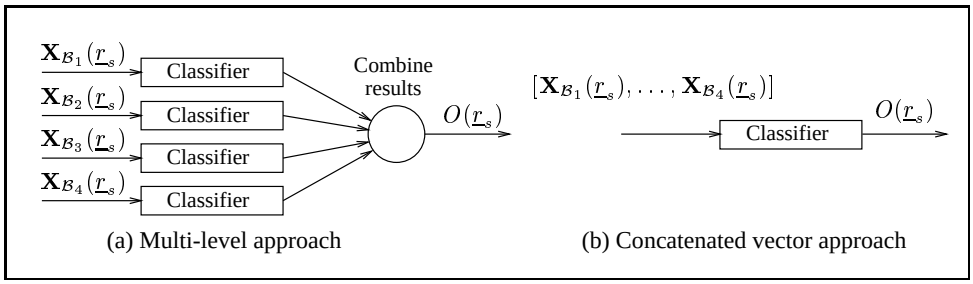

Figure 2: **Proposed architectures.**(a) MLA : Multi-level architecture. (b) CVA : Concatenated vector architecture.

The critically sampled nature of the CWT implies that this integration amounts to summing all samples from scale s and orientation θ that occur in block $\mathcal{B}_l$. In this paper 24 texture components are extracted, corresponding to 6 orientations at 4 scales of wavelet decomposition. Using the terms defined above, the feature vector at a particular block $\mathcal{B}_l$ is found by concatenating the texture based features with the colour based features in the follwoing manner: $\mathbf{X}_{\mathcal{B}_l} = [\mathcal{T}_1^1(\mathcal{B}_l),\ \mathcal{T}_1^2(\mathcal{B}_l),\ \ldots, \mathcal{T}_4^6(\mathcal{B}_l),\ \mathcal{C}_1(\mathcal{B}_l),\ \mathcal{C}_2(\mathcal{B}_l)]$

2.2 Classifiers

The configurations of the two architectures analysed in this paper are illustrated in Figure 2. The motivation behind these configurations is to obtain a robust classification method by aggregating information obtained from a number of levels (block-sizes). Two methods are proposed, the first 'late aggregation' approach, termed henceforth the multi-level architecture 'MLA', combines classification results from separate levels to form an overall robust classification. The second 'early aggregation' approach, termed the concatenated vector architecture 'CVA', concatenates feature vectors and then performs a single classification on the resulting feature vector.

The training data for the classifiers was extracted from 110 natural and 110 man-made homogeneous images.[2] When sampled at 16×16 pixel intervals this gives approximately 60,000 feature vectors at each level in the MLA method and 60,000 feature vectors in total in the CVA method. A validation set of 120 inhomogeneous images and a test set of 120 inhomogeneous images (120 images correspond to 35,000 examples) was used to allow optimisation and testing of the various classifiers and architectures. This validation and test data consisted of 16×16 pixel blocks, each one handlabelled as either natural or man-made.

The classification techniques implemented in this paper are Fisher's linear discriminant (evaluated as a benchmark) termed henceforth FLD, the recently proposed kernel linear discriminant KFD [3] [8], and the support vector method, SVM [11], [15]. This latter method was chosen based on its excellent classification performance in a number

[2]In this paper, the term 'homogeneous' refers to scenes that only contain one class of image data (i.e. in the natural/man-made case the image consists of completely natural or completely man-made objects) whereas 'inhomogeneous' refers to scenes containing both classes.

of fields [5], [11]. Gaussian RBF and homogenous polynomial kernels were analysed and, in the SVM case, so was the straightforward linear approach. An advantage of the discriminant methods is that class conditional probability distributions can be estimated by projecting the training data from each of the classes onto the discriminating vector. Gaussian density models (justified with appeal to the central limit theorem) can then be fitted to these distributions using the standard maximum-likelihood approach (see Section 3).

As is normal in the kernel methods community the dimensions of the feature vectors were rescaled to [-1,+1] to avoid scaling problems when using polynomial kernels. [3]

2.3 Computational issues

Before evaluating the proposed algorithms, the pragmatics of training and testing the classifiers and architectures must be considered. In the simple FLD approach all the training data may be used to evaluate the discriminating vector and the resulting algorithm is very fast in the testing phase.

The KFD method requires the inversion of an $N \times N$ matrix where N is the number of training points under consideration. The choice of N also directly affects the testing phase as this determines the size of the resulting kernel expansion. It is clearly not possible to use the entire training set and thus a regularly spaced subset of data points was selected. For the results in this paper $N = 900$ was chosen, this giving rise to reasonably fast training and testing phases.

The SVM method was trained using the sequential minimal optimisation method developed by Platt, [6]. Unfortunately, the nature of this problem, where the classes are significantly overlapped and the feature vectors consist of real valued (not binary) numbers, implies that using the entire training set to train one classifier is computationally too expensive. Because of this, a similar sub-sampling of the training data to that used in the KFD method was undertaken, in this case using 6000 data points.

Reduced set methods [12] may reduce the testing time complexity for both the KFD and SVM methods. The results presented in this paper are preliminary in nature and this aspect of the work is to be investigated in the future.

2.4 Combining results in the multi-level architecture

For the multi-level approach, results from a number of block-sizes are combined to form a single classification result. To achieve this when using the SVM method a simple voting scheme is used such that each SVM classifier corresponding to each block-size contributes one vote to the decision. In the event of a tie the result from the classifier corresponding to the largest block-size is used.

To combine the results from the class conditional densities when using either of the discriminant methods a naive Bayes classifer approach is used. The assumption of this model is that the likelihoods at each level are statistically independent of each other, an assumption motivated by the need for low computational complexity. Given that feature vectors have been extracted from a number of block-sizes, 1 to L, and the condi-

[3] Note that linear discriminant methods are invariant to this type of rescaling.

a. ρ = 0.92 b. ρ = 0.36 c. ρ = 0.18

Figure 3: **Examples of Natural/Man-made Results.** Top row: Original images. Bottom row: The brightness of each image block is weighted with the posterior probability of being man-made, $P(C_M|\mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4)$. ρ : Proportion of samples classified as man-made.

tional probabilities corresponding to class k: $P(\mathbf{X}_1|C_k), P(\mathbf{X}_2|C_k), \ldots, P(\mathbf{X}_L|C_k)$ have been determined then, based on the previously stated assumption, the posterior conditioned on the data can be evaluated as follows:

$$P(C_k|\mathbf{X}_1, \ldots, \mathbf{X}_L) = \frac{\prod_{l=1}^{L} P(\mathbf{X}_l|C_k)}{P(\mathbf{X}_1, \ldots, \mathbf{X}_L)} P(C_k) \tag{3}$$

If a classifier has a block that reaches beyond the edge of the image, it does not participate in any voting or result combination.

3 Results

Results, using the optimum KFD method, are illustrated in Figure 3. Denoting the man-made class as C_M, the 16×16 pixel area surrounding each sample point has been weighted with the posterior probability, $P(C_M|\mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4)$. These results clearly illustrate the power of the proposed algorithms.

The results obtained from the test set are given in Table 1. The first result to note is that all the MLA approaches perform better than their CVA counterparts. We conjecture that the superiority of the MLA approach is partly caused by the fact that the discrimination tasks vary in difficulty as we move across the levels (as illustrated in Figure 4). The CVA approach has to accommodate all tasks at once, making it more difficult to find a model complexity that suits the structure in the data in the sense that it can identify all regularities without overfitting. In addition, the CVA approach works in a much higher-dimensional feature space, while the training set size is kept constant,

which increases the statistical complexity of the learning task. The original reason for proposing the CVA approach was that we expected that there might be useful information in the interaction between different levels. Our results show that for the considered training set sizes (which were fairly small) this information was not sufficiently useful to offset the above effects.

Architecture	Error (%)	Train	Test	Kernel	C	σ	D	Structure
CVA-Linear-SVM	20.6	27	14					$2, 3, 4$
CVA-Nonlinear-SVM	18.8	50	42	RBF	1.0	6.2		$2, 3, 4$
CVA-KFD	22.3	34	21	Poly			14	$2, 3, 4$
CVA-FLD	24.8	25	4					$2, 3, 4$
MLA-Linear-SVM	19.9	2	42					$2, 3, 4$
MLA-Nonlinear-SVM	17.5	17	66	RBF	1.0	0.07		$2, 3$
MLA-KFD	18.2	5	14	RBF		0.9		$2, 3, 4$
MLA-FLD	23.4	1	5					2

Table 1: **Classification error for all 16×16 pixel blocks extracted from the test set of images.** Also included are the training and testing times in minutes for all architectures. The remaining columns list the optimum choice of kernel, parameters and multi-level structure. The test set consisted of 35000 examples extracted from 120 images.

The best overall result was an error of 17.5% obtained when using a non-linear SVM with an RBF kernel operating over levels 2 and 3 in the hierarchy. However, this is also the most computationally costly. The optimum KFD approach uses an RBF kernel operating over levels 2,3 and 4 in the hierarchy giving an error of 18.2%. The McNemar test [7] was used to determine whether the kernel methods give significantly better results than the linear method. This was confirmed at a confidence level of 0.01.

As described earlier, the conditional densities for each class at each level in the discriminant methods are estimated using Gaussian density models, examples of which are shown in Figure 4. This model is chosen based on the assumption that the projection onto the discriminating vector approximates to a sum of independent random variables thus allowing us to invoke the central limit theorem. Of particular interest is the observation that this approximation is extremely accurate when using the KFD method because of the high dimensionality of the feature space, thus implying that the resulting density estimates are reliable. This makes the KFD approach a compelling choice of algorithm in situations where accurate conditional densities are required.

4 Summary

This paper has analysed a number of architectures and a number of classfiers and found that localised semantic image classification can be performed to a high degree of accuracy. SVMs are based on some fairly advanced methods of capacity control and statistical learning theory, and have achieved record results on a number of benchmarks

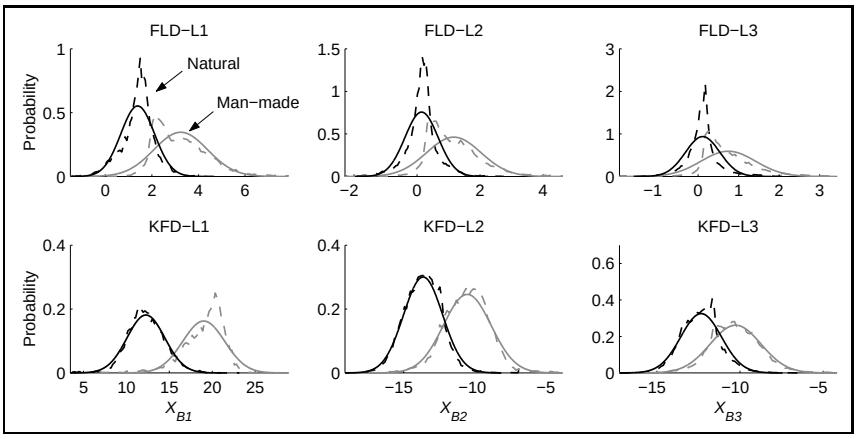

Figure 4: **Estimates of the class likelihood probability densities**. Class likelihoods (solid) and underlying histograms (dashed). Black lines correspond to the 'natural' class, grey lines to the 'man-made' class. Top row: Levels 1, 2 and 3 from the FLD approach. Bottom row: Levels 1, 2 and 3 from the KFD approach.

[11]. Seen in this light, it might be surprising that Fisher's linear discriminant, when carried out in feature space, performs as well as indicated.

All learning algorithms have strengths and weaknesses, related to the implicit assumptions made about the data. One assumption of Fisher's discriminant is that the data are normally distributed, our experiments indicate that this assumption holds true for the data we considered. This, we believe, explains the good performance. In addition, it allows us to use the outputs of the system in subsequent probabilistic inference tasks. These two features, we believe, make a fairly strong case for using KFD in image classification tasks.

Acknowledgements: I would like to thank Mike Tipping and Andrew Blake for useful suggestions relating to this paper.

References

[1] O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram image classification. *IEEE Trans on Neural Networks*, 10(5):1055–1064, September 1999.

[2] N. G. Kingsbury. The dual-tree complex wavelet transform: A new efficient tool for image restoration and enhancement. In *EUSIPCO'98*, volume 1, pages 319–322. EURASIP, 1998.

[3] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing IX*, pages 41–48. IEEE, 1999.

[4] Y. Ohta, T. Kanade, and T. Sakai. Colour information for region segmentation. *Computer Graphics and Image Processing*, 13:222–241, 1980.

[5] C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *ICCV'98*, Bombay, India, January 1998.

[6] J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 185–208, Cambridge, MA, 1999. MIT Press.

[7] B.D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[8] V. Roth and V. Steinhage. Nonlinear discriminant analysis using kernel functions. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*. MIT Press, 2000.

[9] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 20(1), January 1998.

[10] Y. Rubner and C. Tomasi. Texture-based image retrieval without segmentation. In *ICCV'99*, Corfu, Greece, September 1999.

[11] B. Schölkopf, C. J. C. Burges, and A. J. Smola. *Advances in Kernel Methods — Support Vector Learning*. MIT Press, Cambridge, MA, 1999.

[12] B. Schölkopf, S. Mika, C. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. Smola. Input space vs. feature space in kernel-based methods. *IEEE Transactions on Neural Networks*, 10(5):1000 – 1017, 1999.

[13] A. B. Torralba and A. Oliva. Semantic organisation of scenes using discriminant structural templates. In *ICCV'99*, Corfu, Greece, September 1999.

[14] A. Vailaya, M. Figueiredo, A. K. Jain, and H.J. Zhang. Content-based hierarchical classification of vacation images. In *IEEE Conf. on Multimedia Computing and Systems*, volume 1, 1999.

[15] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, N.Y., 1995.