Visual Category Recognition Using Spectral Regression and Kernel Discriminant Analysis

M. A. Tahir, J. Kittler, K. Mikolajczyk, F. Yan
Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, GU2 7XH, UK
{m.tahir,j.kittler,k.mikolajczyk,f.yan}@surrey.ac.uk

K.E.A. van de Sande, T. Gevers
ISLA, Informatics Institute
University of Amsterdam
Science Park 107, Amsterdam, NL
{ksande,gevers}@science.uva.nl

Abstract

Visual category recognition (VCR) is one of the most important tasks in image and video indexing. Spectral methods have recently emerged as a powerful tool for dimensionality reduction and manifold learning. Recently, Spectral Regression combined with Kernel Discriminant Analysis (SR-KDA) has been successful in many classification problems. In this paper, we adopt this solution to VCR and demonstrate its advantages over existing methods both in terms of speed and accuracy. The distinctiveness of this method is assessed experimentally using an image and a video benchmark: the PASCAL VOC Challenge 08 and the Mediamill Challenge. From the experimental results, it can be derived that SR-KDA consistently yields significant performance gains when compared with the state-of-the art methods. The other strong point of using SR-KDA is that the time complexity scales linearly with respect to the number of concepts and the main computational complexity is independent of the number of categories.

1. Introduction

Visual category recognition also referred to as concept detection aims at annotating videos and images using a vocabulary defined by a set of concepts of interest including scenes types (outdoor, vegetation etc), objects (airplane, car etc), events (people-marching etc) and certain named entities (person, place etc). A standard approach to visual category recognition has been established in the community. This approach involves local descriptor computation, vector quantisation via clustering, structured scene or object representation via localised histograms of vector codes, similarity measure for kernel construction and classifier learning. A significant effort has been invested in searching for better solutions in each of these topics. The development of invariant feature detectors and descriptors [16, 20] has provided a means to deal with occlusion, background clutter and geometric image transformation. An orderless collection of such descriptors already provides a robust and discriminative image representation with good recognition results [4]. Such representation can further be improved by optimizing vector quantisation and codebook [29] which allows to reduce the complexity and to equalise the cardinality of image representation. It can also be done on multiple quantisation levels [9]. For well structured objects and scenes, weak geometric relations encoded in spatial location histograms proved successful [14]. This image representation in the form of high dimensional histograms is often used to compute the similarity between images. While there are some variations in the above discussed parts of this recognition approach, for the final classification stage Support Vector Machines have consistently been used since learning robust concept detectors from large-scale visual codebooks is typically achieved by a kernel-based learning model [24, 8, 9, 14, 17, 27, 28, 29]. Other approaches such as Bayesian [4, 6], LDA [15] or AdaBoost [21] have been considered but it has been observed in PASCAL Visual Object Classification challenges [5] and the TRECVid evaluation campaign [23] over the last few years that SVM dominates in terms of image recognition performance.

Linear Discriminant Analysis (LDA) [7], which is one of the most widely used statistical methods, has been proven successful in many classification including face recognition problems. Kernel Fisher Discriminant Analysis [18] and Generalized Discriminant Analysis [1] are the most popular kernel-based extensions of LDA. The main issue however with these approaches is the singularity and the complexity of eigen-value decomposition, in particular for large datasets in image or video retrieval. Regularization techniques [1] or Generalized singular value decomposition [10] can handle singularities while greedy approximation [19] or QR decomposition [30] can speed-up eigen-decomposition.

Spectral methods have emerged as a powerful tool for dimensionality reduction and manifold learning [3]. Recently, Spectral Regression combined with Kernel Discriminant Analysis (SR-KDA) introduced by Cai et al [2] has been successful in many classification tasks such as multi-class face, text and spoken letter recognition. The method combines the spectral graph analysis and regression for an efficient large matrix decomposition in KDA. It has been demonstrated in [2] that it can achieve an order of magnitude speedup over the eigen-decomposition while producing smaller error rate compared to state-of-the-art classifiers.

Mathematically, the visual category recognition problem can be formulated as a two class pattern recognition problem. The original data set is divided into N data sets where $Y = \{1, 2, ..., N\}$ is the finite set of concepts. The task is to learn one binary classifier $h_a : X \rightarrow \{\neg a, a\}$ for each concept $a \in Y$. In many practical situations, the number of concepts can be very high and learning of the independent binary classification tasks may become computationally expensive especially if kernel-based learning model is adopted. The aim of this paper is to investigate the effectiveness of SR-KDA for large scale visual category recognition since by using SR-KDA in binary classification tasks, the time complexity scales linearly with respect to N. The main computationally intensive operation is Cholesky decomposition, which is actually independent of N. Further, we show that it can produce classification results superior to existing approaches. This makes the proposed solution very convenient as it can directly replace the currently favoured learning method such as SVM without changing the other parts of the system. We extensively test its recognition performance on the challenging Pascal 2008 dataset which consists of 20 object categories and the Mediamill data set comprising 101 concepts. The evaluation of SR-KDA on these data sets enables us to compare its performance with many state-of-the-art approaches. The results clearly indicate its advantage over other approaches. Overall, the presented approach has highest average precision in 12 out of 20 categories for Pascal VOC 2008 and in 57 out of 101 categories. The median average precision is better than all other methods both in Pascal VOC 2008 and Mediamill Challenge.

This paper is organised as follows. Section 2 discusses kernel discriminant analysis using spectral regression along with complexity analysis followed by VCR using SR-KDA in Section 3. Experiments are discussed in Section 4 followed by the results and discussion in Section 5. Section 6 concludes the paper.

2. Kernel Discriminant Analysis using Spectral Regression (SR-KDA)

Kernel Discriminant Analysis is a nonlinear extension of LDA which maps the original measurements into a higher dimensional space using the "kernel trick". Let $\mathbf{x}_i$ be training vectors $\mathbf{x}_i \in \mathcal{R}^d, i = 1, \cdots, m$. K is an $m \times m$ kernel matrix. If ν denotes a projective function into the kernel feature space, then the objective function for KDA is

$$\max_{\nu} \quad D(\nu) = \frac{\nu^T C_b \nu}{\nu^T C_t \nu} \qquad (1)$$

where C_b and C_t denote the between-class and total scatter matrices in the feature space respectively. Equation 1 can be solved by the eigen-problem $C_b = \lambda C_t$. It is proved in [1] that equation 1 is equivalent to

$$\max_{\alpha} \quad D(\alpha) = \frac{\alpha^T KWK\alpha}{\alpha^T KK\alpha} \qquad (2)$$

where $\alpha = [\alpha_1, \alpha_2, ...\alpha_m]^T$ is the eigen-vector satisfying $KWK\alpha = \lambda KK\alpha$. $W = (W_l)_{l=1,....n}$ is a $(m \times m)$ block diagonal matrix of labels arranged such that upper block corresponds to positive examples and lower one to negative examples of the class. Each eigenvector α gives a projection function ν in the feature space.

It is shown in [2] that instead of solving the eigen-problem in equation 2, the KDA projections can be obtained by the following two linear equations

$$\begin{aligned} W\phi &= \lambda\phi \\ (K + \delta I)\alpha &= \phi \end{aligned} \qquad (3)$$

where ϕ is an eigenvector of W, I is the identity matrix and $\delta > 0$ is a regularisation parameter. Eigen-vectors ϕ are obtained directly from Gram-Schmidt method. Since $(K + \delta I)$ is positive definite, the Cholesky decomposition is used to solve the linear equations in equation 3. Thus, SR-KDA only needs to solve a set of regularised regression problems [2] and there is no eigenvector computation involved. This results in great improvement of computational cost and allows to handle large kernel matrices.

2.1. Complexity Analysis

The computation of SR-KDA involves two steps: (i) response generation which is the cost of the Gram-Schmidt method (ii) regularised regression which involves solving $(c - 1)$ linear equations using the Cholesky decomposition where c is the number of classes. As in [26], we use the term flam, a compound operation consisting of one addition and one multiplication, to measure the operation counts. The cost of the Gram-Schmidt method requires $(mc^2 - \frac{1}{3}c^3)$ flams. The Cholesky decomposition requires $\frac{1}{6}m^3$ flams and the $c - 1$ linear equations can be solved within m^2c

flams. Thus, the computational cost of SR-KDA excluding the cost of Kernel Matrix K is $\frac{1}{6}m^3 + m^2c + mc^2 - \frac{1}{3}c^3$ which can be approximated as [2]

$$\frac{1}{6}m^3 + m^2c$$

Comparing to the cost of ordinary KDA ($\frac{9}{2}m^3 + m^2c$) [2], SR-KDA significantly reduces the dominant part and achieves a 27 times speedup.

3. Visual Category Recognition using SR-KDA

As discussed in Section 1, the visual category recognition problem can be formulated as a two class pattern recognition problem. The original data set is divided into N data sets where $Y = \{1, 2, ..., N\}$ is the finite set of concepts. The task is to learn one binary classifier $h_a : X \rightarrow \{\neg a, a\}$ for each concept $a \in Y$. Among the advantages of using SR-KDA in binary classification tasks is that its time complexity scales linearly with respect to N. The total computational cost of SR-KDA for all concepts is

$$\frac{1}{6}m^3 + m^2Nc$$

The above analysis clearly shows the effectiveness of SR-KDA for visual category recognition. Figure 1 shows the traditional visual category recognition system in which separate classifier is needed to learn individual concept. Figure 2 shows the approach in which cholesky decomposition is first performed only once irrespective of the number of the concepts and then only N linear equations are solved that results in significant reduction in computational cost. In summary, using proposed approach, the main task is to perform Cholesky decomposition and then to solve N linear equations which requires only m^2Nc flams.

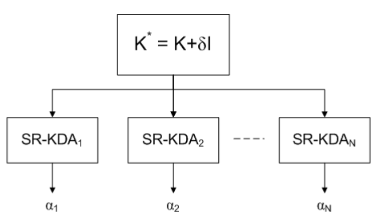

Figure 1. Traditional Visual Category Recognition System.

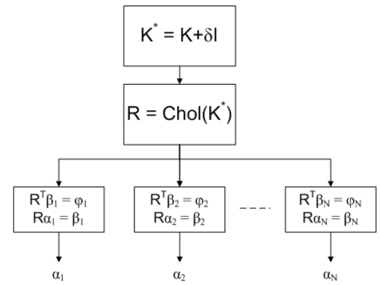

Figure 2. Visual Category Recognition using SR-KDA. The most computational part i.e. Cholesky decomposition is performed only once irrespective of the number of concepts.

4. Experimental Setup

In this section we give the implementation details of our recognition system, discuss the training and test data as well as the experimental settings.

4.1. Datasets

The PASCAL Visual Object Classes Challenge [5] provides a yearly benchmark for comparison of object classification methods. The The Pascal VOC 2008 dataset consists of 8465 images of 20 different object classes such as aeroplane, bus, cat, etc. The dataset is divided into a predefined "trainval" set (4332 images) and "test" set (4133 images). The "trainval" dataset is further divided for validation purpose into a training set containing 2111 images and a validation set containing 2221 images. The ground truth for test sets is not released to avoid over-fitting of classifiers. The objective of the classification tasks is to make 20 binary decisions for each given test image as to whether it contains any of the 20 objects. The results are then evaluated independently on test set by the Pascal organisers. To give more insight into the evaluated methods some of the proposed solutions are tested on the validation set.

Video Benchmark (Mediamill Challenge) The Mediamill Challenge provides an annotated video dataset, based on the training set of NIST TRECVID 2005 benchmark [23]. This dataset consists of 86 hours of video, divided into a challenge training set (70% of the data or 30993 examples) and test set (30% the data or 12914 examples). The dataset content comprises television news from November 2004 broadcasted on six different TV channels.

4.2. Evaluation Criteria

The average precision is a single-valued measure that is proportional to the area under a precision-recall curve. This value is the average of the precision over all relevant judged shots. This metric combines precision and recall into one performance value. This measure is computed from the ranking list of all the key frames in the database established by ordering their similarities to a specified concept. Average Precision for each concept (AP) is defined as

$$AP = \frac{1}{|R|} \sum_{k=1}^{|R|} c_k \tag{4}$$

where R is the number of positive samples in a test set and the contribution c_k of the k^{th} element in the ranking list is defined as

$$c_k = \{ \begin{array}{lll} \frac{|R \cap M_k|}{k} & if & concept\ true \\ 0 & if & concept\ not\ true \end{array} \tag{5}$$

where $M_k = \{i_1, i_2, ..., i_k\}$ is a ranked list of the top k retrieved samples from the test set.

For Pascal VOC data set average precision is calculated independently by the Pascal VOC Organizers over 11 thresholds on recall $r\{0, 0.1, ..., 0.9, 1\}$ for which the interpolated precision p(r) is computed and then the arithmetic mean is taken.

4.3. Image Representation

The bag-of-words model [22] has become the method of choice for visual categorisation [5, 28, 8, 31]. This model first extracts specific points in an image using a point sampling strategy. Descriptors are computed for the neighbourhood of these points, which represent the local area. The bag-of-words model performs vector quantisation of the descriptors in an image against a visual codebook. A descriptor is assigned to those codebook elements which are closest in Euclidean space using soft assignment [29]. This results in a fixed-length representation of the image.

As a point sampling strategy, two methods have been chosen: dense sampling and Harris-Laplace salient points. Dense sampling samples points regularly over the image at fixed pixel intervals. Typically, around 10,000 points are sampled per image at an interval of 6 pixels. The Harris-Laplace salient point detector [20] uses the Harris corner detector to find potential feature locations and then selects a subset of these points for which the Laplacian-of-Gaussians reaches a maximum over scale.

To describe the area around the sampled points, we use the SIFT descriptor [16] and four extensions of SIFT to colour [28]: OpponentSIFT, rgSIFT, C-SIFT and RGB-SIFT. These descriptors have specific invariance properties with respect to common changes in illumination conditions and have been shown to improve visual categorisation accuracy [28]. To construct the visual codebook, descriptors are clustered within each type using k-means and form codebooks of 4000 clusters each.

The image is divided into 3 spatial location grids: entire image, image quarters (2×2) and horizontal bars (1×3). It is then represented by histograms of codebook occurrences in the spatial grids [14].

In the next two subsections, the usage of these visual features for the Pascal VOC 2008 and the Mediamill Challenge is discussed. It should be noted that in demonstrate the benefit of using SR-KDA we build images representations similar to those used by other systems previously tested on these datasets and in which SVM is used as the classifier.

Pascal VOC 2008 To obtain visual features for the Pascal VOC 2008 dataset, we combine the 2 sampling strategies, 5 different descriptor types and 3 spatial location grids listed in the previous section. This results in 30 different visual feature representations per image. For each visual feature, a separate kernel matrix is created. The kernel function used to compute entries in the kernel matrix are based on the χ^2 distance for feature vectors $\vec{F}$:

$$k(\vec{F}, \vec{F'}) = e^{-\frac{1}{A} dist_{\chi^2}(\vec{F}, \vec{F'})} \tag{6}$$

where A is a scalar which normalises the distances. Following [31], A needs to be set to the average χ^2 distance between all elements of the kernel matrix.

Classification is performed by both classifier-level (SR-KDA$_1$) and kernel-level fusion (SR-KDA$_2$). In classifier-level fusion, 30 classifiers are first learnt using each kernel independently and then classifiers score are fused. Two simple voting rules for classifiers combination namely $\{AVG, MAX\}$ [12] are adopted. In kernel-level fusion, first all kernels are combined using equal weights and then this combined kernel is used for classification.

Mediamill Challenge For Mediamill Challenge, two different kernel matrices are used. The first kernel matrix is constructed from the publicly available baseline features of the Mediamill Challenge [24]. This baseline feature has a length of 120. The second kernel matrix, use the OpponentSIFT descriptor which gives the best performance in [28] and consists of Harris-Laplace and dense sampling strategies and the spatial location grids for entire image and the image quarters.

5. Results and Discussion

This section presents a number of experiments carried out on the data and according to the criteria defined in the previous section. It is important to note that the results for the compared methods are directly taken from [5] for Pascal VOC 2008 and [24, 27] for Mediamill Challenge. The regularisation parameter $\delta = 0.01$ is used in all experiments and is chosen after experiments on the validation set of Pascal VOC 2008.

5.1. Results on Image Benchmark (Pascal VOC 2008)

We first evaluate classifiers performance on the Pascal VOC 2008 validation set and then compare to the state-of-the art systems that produced top results in Pascal Challenge. Table 1 presents the scores for classifier-level and kernel-level fusion for SR-KDA and SVM. The AVG rule for combining classifier outputs of SR-KDA$_1$ gives consistently better results than using the MAX operator. We therefore adopt the AVG strategy for further experiments. The results show that SR-KDA performs better than SVM in most of the categories. There is an increase of 2.5% in Median AP when SR-KDA$_1$ is compared with SVM$_1$ (classifier-level fusion) and 2.4% when SR-KDA$_2$ is compared with SVM$_2$ (kernel-level fusion). The results indicate a performance increase when kernel weights are learned using multi-kernel SVM (MKSVM) [25, 13]. Future work aims to investigate methods such as multi-kernel KDA [11] to learn kernel weights in the presented approach since MKSVM performs better than equal weights SVM (SVM_2). Overall, both $SR\text{-}KDA_1$ and $SR\text{-}KDA_2$ have the highest performance in 8 out of 20 concepts respectively, followed by multi-kernel SVM (MKSVM), SVM_1 and SVM_2 in 3, 2 and 2 categories respectively.

We compare our results with top methods submitted in Pascal VOC 2008 [5] as shown in Table 2. Technical details of these approaches have not been published but from personal communication we infer that L-Shotgun and L-Flat is based on approach from [8], U-TreeSFS and U-Soft5ColorSift are variants of [28]. In all these methods, SVM is used at learning stage and the multi-class is addressed in a one-vs-all set-up. It is clear that by using SR-KDA either at classifier-level or kernel-level, the Median AP is better than for all other methods. The results show that the classifier-level combination of $SR\text{-}KDA_1$ performs best in terms of Median AP. $SR\text{-}KDA_1$ ranked first in Bottle, Bus, Chair, Cow, and Train while $SR\text{-}KDA_2$ ranked first in Bicycle, Bird, Car, Horse, Person, PottedPlant, and Sheep. In total the SR-KDA based classifiers perform best in 12 out of 20 concepts. Overall, there is a improvement of 1.8% when compared with the best method in Pascal VOC 2008 competition.

5.2. Results on Video Benchmark (Mediamill Challenge)

Table 3 shows the average precision (AP) for each category in Mediamill data set using presented method SR-KDA, and SVM [27, 24]. It is clear from this table that significant improvement (approx. 51%) is achieved when SR-KDA classifier is compared with SVM using publicly available baseline feature vector. Overall, SR-KDA has highest performance in 57 out of 101 categories when OpponentSIFT image representation is used while SVM has highest performance in 49 categories. The results indicate that SR-KDA performs quite well when class imbalance is not severe (for example in People, Face, Crowd etc). On the other hand, in many highly unbalanced categories like Basketball, Cartoon etc., SVM performs well. Overall, SR-KDA consistently performs better than SVM in most of the concepts and there is an improvement of 3% when compared with the the state-of-the-art method for Mediamill Challenge [27].

5.3. Execution Time

Table 4 shows the training time for SR-KDA, and SVM using a pre-computed kernel for Pascal VOC 2008 (trainval, 2111×2111 kernel matrix) and Mediamill Challenge (30993×30993 kernel matrix). All the experiments have been performed on a $16 \times 3GHz$ hyperthreaded CPUs and 128 GBytes of memory. We used C++ implementation for SVM from the publicly available machine learning toolbox SHOGUN[1] and Matlab implementation for KDA, but the crucial operations (Cholesky decomposition) are in native language. For category recognition, one binary classifier is needed for each concept. For Pascal VOC 2008, there are 20 categories and SR-KDA trains all categories in just 3.3 seconds including 0.66 seconds for Cholesky decomposition while SVM requires 16.4 seconds to train. For Mediamill challenge, there are 101 categories and thus 101 binary classifiers to train which require only 1.3 hours using SR-KDA. SVM requires on average 18 minutes to train each category and in total 30.3 hours. This time complexity analysis clearly indicates the effectiveness of SR-KDA in learning stage for large scale visual category recognition.

Data set		SR-KDA	SVM
Pascal VOC 2008	Decomposition	$0.66s$	-
	Training (20 classifiers)	$2.6s$	$16.37s$
	Total	$3.26s$	$16.37s$
Mediamill	Decomposition	$35m$	-
	Training (101 classifiers)	$51m$	$30.3h$
	Total	$1.26h$	$30.3h$

Table 4. Execution Time for SR-KDA and SVM during training. h = Hours, m = Minutes and s = seconds.

5.4. Discussion

The results show the usefulness of kernel discriminant analysis using spectral regression at the learning stage for visual category recognition. SR-KDA avoids expensive eigen-value decomposition and thus it is possible to evaluate the performance of KDA on large scale experiments especially on the Mediamill data set which consists of 30993 images. For visual category recognition, the SR-KDA algorithm showed to be significantly faster than SVM especially for large number of concepts while simultaneously leading to improvement in the classification performance in most of the categories. SR-KDA also inherits the convenient property of data visualisation, since it allows low dimensional views of the data vectors. This makes an intuitive analysis possible, which is helpful in many practical applications. In summary, considering both accuracy and efficiency, the presented approach is the best choice among the compared approaches. It provides an efficient and effective learning solution for large scale data sets.

Figure 3 and 4 shows the top 16 images retrieved by the presented system for bicycle and motobike respectively using Pascal VOC validation set. This illustrates the intra class variability of appearances as well as interclass similarities that can be correctly classified. For bicycle, 5 images are wrongly identified by the system (Images 3, 5, 10, 11,

[1] http://www.shogun-toolbox.org/

Concept	Train	Validate	SVM_1	SVM_2	MKSVM	$SR\text{-}KDA_1$		$SR\text{-}KDA_2$
						AVG	MAX	
Aeroplane	119	117	75.2	74.6	**79.5**	79.2	78.7	79.3
Bicycle	92	100	37.8	38.1	38.1	39.7	37.5	**40.2**
Bird	166	139	47.7	48.9	51.5	49.1	42.6	**52.1**
Boat	111	96	60.0	59.0	63.2	62.6	57.3	**63.3**
Bottle	129	114	18.2	17.4	17.6	**18.8**	16.2	18.3
Bus	48	52	49.4	53.4	53.0	52.3	47.9	**55.1**
Car	243	223	54.4	53.8	53.9	**55.5**	50.1	54.4
Cat	159	169	55.1	53.8	54.9	**56.2**	50.8	55.5
Chair	177	174	42.1	41.4	41.9	**42.5**	39.6	41.3
Cow	37	37	**21.1**	15.5	17.1	24.1	17.8	18.8
Diningtable	53	52	24.5	24.3	**25.8**	25.7	25.5	26.6
Dog	186	202	**35.0**	34.0	32.6	32.8	30.4	32.5
Horse	96	102	45.3	45.1	**47.1**	**47.1**	42.5	**47.1**
Motorbike	102	102	37.6	39.1	40.1	39.9	34.2	**41.5**
Person	947	1055	85.8	86.3	88.5	88.0	86.9	**88.5**
Pottedplant	85	95	22.5	**25.8**	25.7	24.1	16.6	**25.8**
Sheep	32	32	30.1	29.8	29.9	**30.8**	25.7	29.8
Sofa	69	65	38.4	34.5	36.9	**38.7**	29.0	35.7
Train	78	73	63.6	64.1	65.4	**68.5**	61.8	67.6
Tvmonitor	107	108	51.0	**53.7**	53.0	50.9	47.9	53.6
Median AP			43.7	43.3	44.5	**44.8**	41.1	44.3

Table 1. Classifiers performance on the Pascal VOC 2008 validation set.

Concepts	Train	Test	T-SBFS	X-RCE	U-Flat	L-Soft5	U-TreeSFS	L-Shotgun	$SR\text{-}KDA_1$	$SR\text{-}KDA_2$
			Top methods in Pascal VOC 2008 Challenge							
Aeroplane	236	236	77.9	78.9	80.1	79.7	80.8	**81.1**	79.5	79.8
Bicycle	192	187	47.3	48.0	51.8	52.1	53.2	52.9	54.3	**54.6**
Bird	305	304	52.4	58.7	60.5	61.5	61.6	61.6	61.4	**62.4**
Boat	207	209	61.0	65.2	66.9	65.5	65.6	**67.8**	64.8	65.8
Bottle	243	243	27.9	29.0	29.1	29.1	29.4	29.4	**30.0**	29.5
Bus	100	90	45.5	44.8	52	46.5	49.9	**52.1**	**52.1**	49.4
Car	466	466	53.5	56.1	57.4	58.3	58.5	58.7	59.5	**59.6**
Cat	328	331	55.5	56.3	58.6	57.4	59.4	**59.9**	59.4	58.9
Chair	351	349	47.6	43.7	48.7	48.2	48	48.5	**48.9**	48.8
Cow	74	76	26.8	32.8	31.0	27.9	30.1	32.0	**33.6**	33.1
DiningTable	105	104	**40.8**	30.4	39.2	38.3	39.6	38.6	37.8	36.6
Dog	388	366	46.1	39.7	47.6	46.6	45.0	**47.9**	46.0	47.3
Horse	198	194	58.6	61.2	64.2	66.0	67.3	65.4	66.1	**67.7**
MotorBike	204	203	58.3	61.7	64.6	60.6	60.4	**65.2**	64.0	62.0
Person	2002	1826	83.5	86.8	87.0	87.0	87.1	87.0	86.8	**87.2**
PottedPlant	180	177	26.4	22.9	28.6	31.8	30.1	29.0	29.2	**33.1**
Sheep	64	66	24.3	34.2	33.3	42.2	41.5	34.4	42.3	**43.4**
Sofa	134	134	39.2	44.2	42.6	45.3	**45.4**	43.1	44.0	43.6
Train	151	151	70.3	68.4	73.1	72.3	74.3	74.3	**77.8**	76.5
TV/Monitor	215	200	56.9	59.1	59.8	**64.7**	59.8	61.5	61.2	63.7
Median AP			50.0	52.1	54.7	54.8	55.9	55.8	**56.9**	56.8

Table 2. Classifiers performance on the Pascal VOC 2008 test set.

12 row-wise in Figure 3) and for motorbike, 3 images are wrongly identified (Images 5, 13, 16 row-wise in Figure 4).

In this paper, for Pascal VOC 2008 data set, equal weights are used in kernel-level fusion for SR-KDA. It is shown in Table 1 that learning weights using multi-kernel SVM (MKSVM) can improve the performance of equal weights SVM. Future work aims at improving performance by investigating methods such as multi-kernel KDA to learn kernel weights for the presented approach.

6. Conclusions

Kernel discriminant analysis using spectral regression (SR-KDA) is introduced in this paper at the learning stage of visual category recognition. The presented method uses regression instead of expensive eigen-value decomposition of the kernel-matrix to solve the optimisation problem. Its recognition performance is evaluated on the challenging Pascal 2008 dataset which consists of 20 object categories and the Mediamill data set comprising 101 concepts. From the experimental results, it can be derived that SR-KDA consistently yields performance gains when compared with the state-of-the art methods and the time complexity scales linearly with respect to the number of concepts. The main computationally expensive operation is Cholesky decomposition which is performed only once irrespective of the number of categories.

Acknowledgements

This work was supported by the EU VIDI-Video Project.

References

[1] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 2(12):2385–

Concept	Train	Test	Baseline		Opponentsift		Concept	Train	Test	Baseline		Opponentsift	
			SVM [24]	SR-KDA	SVM [27]	SR-KDA				SVM [24]	SR-KDA	SVM [27]	SR-KDA
People	77.67	75.87	83.1	94.7	96.7	**97.3**	Table	0.75	0.52	7.3	7.3	**15.1**	12.5
Face	64.15	62.37	89.5	91.2	95.6	**96.2**	Tower	0.75	0.63	5.7	6.5	**25.9**	25.4
Overlayed-text	36.33	34.30	66.9	71.6	86.9	**88.2**	Basketball	0.69	0.34	38.2	30.8	**76.5**	65.2
Outdoor	32.68	38.33	68.8	77.5	86.3	**87.8**	Arrafat	0.62	0.88	2.6	3.3	**30.6**	28.0
Entertainment	19.64	12.55	16.6	37.0	66.9	**71.0**	Chair	0.60	0.58	48.6	50.1	56.7	**59.6**
Indoor	19.59	21.20	59.3	67.6	78.9	**80.3**	Explosion	0.53	1.04	9.8	15.3	**24.1**	22.7
Studio	13.66	14.20	63.6	73.1	87.4	**88.1**	Flag-usa	0.50	0.94	28.7	37.6	51.2	**51.4**
Walking-running	13.61	16.83	35.3	39.1	51.4	**53.7**	Bus	0.43	0.64	1.3	3.7	**5.8**	5.7
Urban	11.78	8.80	22.2	25.5	40.3	**43.0**	Snow	0.41	0.53	8.5	12.6	32.2	**33.8**
Crowd	11.48	16.12	48.0	53.4	65.6	**68.4**	Football	0.35	0.39	12.1	11.1	**73.2**	72.4
Sky	10.77	11.38	47.8	54.5	**71.1**	70.5	Tennis	0.34	0.56	44.8	48.5	**81.7**	78.7
Government-leader	9.35	7.87	21.3	24.1	39.2	**41.5**	Prisoner	0.33	0.22	4.7	15.9	**27.8**	26.8
Violence	8.07	9.75	31.7	37.7	51.2	**53.2**	Newspaper	0.31	0.27	37.5	40.5	**67.5**	63.4
Road	7.76	6.60	19.5	23.2	**39.1**	38.6	Lahoud	0.30	0.15	28.9	38.9	**35.9**	35.8
Vehicle	7.61	8.53	22.1	33.0	49.2	**51.8**	Kerry	0.29	0.01	0.0	0.0	0.0	0.0
Building	6.86	11.16	31.6	33.8	45.5	**48.7**	House	0.29	0.36	2.3	3.7	**9.4**	9.0
Male	5.71	2.38	8.6	12.7	**15.3**	**15.3**	Government-building	0.27	0.19	1.1	3.6	**53.7**	47.1
Anchor	5.09	4.85	63.1	77.5	92.4	**93.3**	Religious-leader	0.27	0.00	4.3	3.1	**20.3**	19.8
Car	4.87	5.93	25.2	29.8	46.6	**49.5**	Fireweapon	0.27	0.52	48.9	64.0	8.9	**12.7**
Meeting	4.53	4.86	25.7	27.5	44.5	**45.8**	Duo-anchor	0.26	0.18	63.4	76.3	**82.2**	80.1
Female	4.38	2.11	8.6	8.9	**22.4**	21.8	Golf	0.25	0.31	9.1	30.9	**35.9**	34.4
Military	4.14	6.58	21.7	23.3	**33.2**	32.7	Allawi	0.21	0.02	0.0	0.0	**0.3**	0.1
Vegetation	3.87	4.64	18.3	23.0	38.7	**41.0**	Bicycle	0.20	0.04	0.6	0.5	73.6	**80.2**
Sports	3.76	2.61	30.4	30.4	**58.3**	57.6	Court	0.20	0.30	9.3	14.3	**54.0**	49.4
Monologue	3.10	2.33	9.4	12.2	39.0	**43.4**	Bush-sr	0.20	0.01	0.0	0.0	0.0	0.0
Graphics	2.89	3.48	36.5	51.4	72.2	**72.8**	Food	0.20	0.83	4.8	5.8	60.7	**60.9**
Corporate-leader	2.57	1.30	1.6	1.7	2.2	**2.3**	Cycling	0.18	0.03	4.2	0.6	95.0	**100.0**
Waterbody	2.31	1.89	15.0	25.2	56.9	**57.8**	Bird	0.18	0.23	72.4	87.4	**93.4**	**93.4**
People-marching	1.93	4.13	22.8	28.5	36.6	**39.4**	Drawing	0.17	0.17	26.5	28.6	**75.0**	61.1
Soccer	1.67	0.29	50.3	47.5	**91.3**	89.0	Horse	0.16	0.02	0.0	0.0	**0.1**	**0.1**
Mountain	1.64	1.01	14.1	24.8	**42.4**	40.5	Dog	0.14	0.38	22.5	31.2	**42.2**	42.0
Bush-jr	1.61	0.54	6.2	4.2	15.8	**16.6**	Nightfire	0.14	0.05	52.6	57.2	**50.8**	50.7
Office	1.56	1.75	7.7	10.4	15.1	**17.3**	Horse-racing	0.12	0.02	0.0	0.0	0.1	**0.2**
Screen	1.53	1.90	10.1	15.2	33.5	**34.6**	River	0.10	0.09	31.0	78.9	**95.4**	**95.4**
Fish	1.26	0.12	18.9	21.2	**89.1**	**89.1**	Racing	0.09	0.12	2.9	1.1	5.1	**5.6**
Truck	1.16	1.02	3.8	4.3	**7.0**	**7.0**	Candle	0.08	0.10	1.1	2.9	**10.1**	9.0
Maps	1.16	1.21	47.6	62.4	**83.7**	82.4	Cartoon	0.08	0.21	25.9	28.7	**40.9**	37.5
Smoke	1.13	2.14	25.0	35.5	**51.1**	49.1	Drawing-cartoon	0.08	0.38	29.3	45.1	**51.3**	48.2
Animal	1.00	0.91	20.9	36.0	**56.5**	56.4	Tank	0.08	0.08	0.8	1.8	**10.3**	**10.3**
Weather	0.99	1.25	40.5	49.1	80.9	**81.3**	Swimmingpool	0.08	0.10	0.3	1.6	18.4	**23.7**
Aircraft	0.99	0.94	7.3	10.2	21.4	**23.8**	Beach	0.08	0.06	2.7	2.3	**7.3**	6.9
Police-security	0.92	0.77	1.2	2.3	**20.3**	19.0	Waterfall	0.07	0.08	38.1	43.4	**60.1**	**60.1**
Flag	0.92	1.12	22.7	21.3	**40.3**	39.7	Motorbike	0.05	0.16	0.6	0.9	**0.2**	**0.2**
Grass	0.90	0.59	6.4	14.2	**34.0**	32.3	Clinton	0.05	0.21	0.4	20.9	59.0	**61.8**
Cloud	0.87	1.54	11.7	15.7	32.5	**33.5**	Tony-blair	0.05	0.26	0.5	0.7	5.4	**6.1**
Splitscreen	0.86	0.60	63.0	70.6	88.1	**88.5**	Hassan-nasrallah	0.05	0.19	**0.6**	**0.6**	0.3	0.3
Desert	0.81	1.44	10.3	13.5	19.3	**19.6**	Powell	0.05	0.47	1.0	0.9	**2.4**	1.3
Natural-disaster	0.81	0.93	5.5	6.4	11.7	**12.8**	Sharon	0.04	0.19	5.0	0.4	**17.9**	13.6
Boat	0.78	0.54	9.6	14.1	41.7	**42.5**	Hu-jintao	0.03	1.03	3.0	2.1	**6.5**	3.6
Tree	0.78	0.84	12.4	13.4	23.9	**25.8**	Baseball	0.01	0.41	0.3	0.5	1.1	**1.7**
Charts	0.76	0.51	32.7	43.9	62.9	**65.2**	**Median AP**			14.1	21.3	40.3	**41.5**

Table 3. Classifiers performance on the Mediamill Challenge.

2404, 2000.

[2] D. Cai, X. He, and J. Han. Efficient kernel discriminat analysis via spectral regression. In *Proceedings of the International Conference on Data Mining*, 2007.

[3] D. Cai, X. He, and J. Han. Spectral regression for efficient regularized subspace learning. In *Proceedings of the 11th International Conference on Computer Vision*, 2007.

[4] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

[5] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. Pascal VOC workshop, 2008.

[6] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594 – 611, 2006.

[7] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[8] A. Gaidon, M. Marszałek, and C. Schmid. The PASCAL visual object classes challenge 2008 submission. Technical report, INRIA-LEAR, 2008.

[9] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, 8:725–760, 2007.

[10] P. Howland and H. Park. Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):995–1006, 2004.

[11] S.-J. Kim, A. Magnani, and S. Boyd. Optimal kernel selection in kernel fisher discriminant analysis. In *Proceedings of the 23rd International Conference on Machine learning*, 2006.

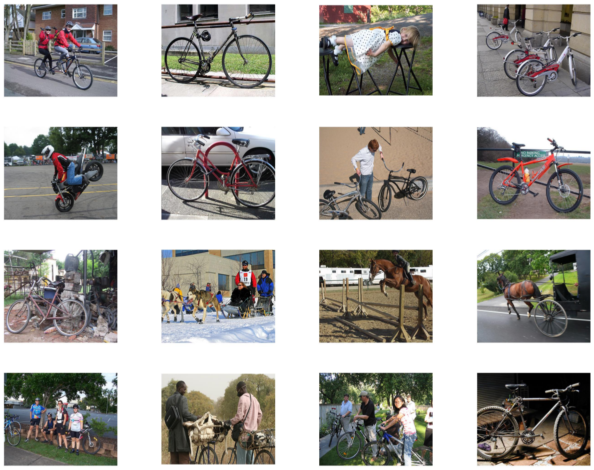

Figure 3. Top 16 images retrieved by the presented system for Bicycle in Pascal VOC Validation Set.

Figure 4. Top 16 images retrieved by the presented system from MotorBike in Pascal VOC Validation Set.

[12] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[13] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

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

[15] X. Liu, L. Zhang, M. Lib, H. Zhang, and D. Wang. Boosting image classification with LDA-based feature combination for digital photograph management. *Pattern Recognition*, 38(6):887–901, 2005.

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60(2):91–110, 2004.

[17] M. Marszałek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition. Pascal Workshop 2007.

[18] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Muller. Fisher discriminant analysis with kernels. In *Proceedings of the Neural Networks for Signal Processing*, 1999.

[19] S. Mika, A. Smola, and B. Scholkopf. An improved training algorithm for kernel fisher discriminants. In *Proceedings of AISTATS*, 2001.

[20] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[21] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer. Generic object recognition with boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):1531–1565, 2006.

[22] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, 2003.

[23] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *ACM International Workshop on Multimedia Information Retrieval*, 2006.

[24] C. G. M. Snoek, M. Worring, J. C. van Gemert, J. M. Geusebroek, and A. W. M. Smeulders. The challenge problem of automated detection of 101 semantic concepts in multimedia. In *Proceedings of the ACM International Conference on Multimedia*, 2006.

[25] S. Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.

[26] G. W. Stewart. *Matrix Algorithms Volume I: Basic Decomposition*. SIAM, 1998.

[27] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. A comparison of color features for visual concept classfication. In *ACM International Conference on Image and Video Retrieval*, 2008.

[28] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluation of color descriptors for object and scene recognition. In *Proceedings of the International Conference on CVPR*, 2008.

[29] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. In *Proceedings of the European Conference on Computer Vision*, 2008.

[30] T. Xiong, J. Ye, Q. Li, V. Cherkassky, and R. Janardan. Efficient kernel discriminant analysis via QR decompostion. *Proceedings of the Advances in Neural Information Processing Systems*, 2004.

[31] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.