Separable Linear Discriminant Classification

Christian Bauckhage and John K. Tsotsos

Centre for Vision Research, York University, Toronto, ON, M3J 1P3
http://cs.yorku.ca/LAAV

Abstract. Linear discriminant analysis is a popular technique in computer vision, machine learning and data mining. It has been successfully applied to various problems, and there are numerous variations of the original approach. This paper introduces the idea of *separable* LDA. Towards the problem of binary classification for visual object recognition, we derive an algorithm for training separable discriminant classifiers. Our approach provides rapid training and runtime behavior and also tackles the small sample size problem. Experimental results show that the method performs robust and allows for online learning.

1 Introduction

Linear discriminant analysis (LDA) is a powerful tool for dimensionality reduction and classification [1,2]. Its applications and extensions are far too numerous to allow for an exhaustive review here. Instead, in this paper, we will restrict our discussion to the linear discriminant analysis of two classes. We shall call the two classes ω_p and ω_n where the subscripts p and n stand for *positive* and *negative*, respectively. Given a set of feature vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L\}$ containing positive and negative examples, binary LDA seeks a projection $\boldsymbol{w}^T\boldsymbol{x}_i$ of the samples that maximizes the inter-class distance of the resulting scalars.

The most widely applied technique for finding the direction $\boldsymbol{w}$ of the optimal projection dates back to seminal work by Fisher [3]. He proposed to determine $\boldsymbol{w}$ by maximizing the Rayleigh quotient $\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}/\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}$ where $\boldsymbol{S}_b$ and $\boldsymbol{S}_w$ are matrices that denote the between-class and within-class scatter of the data. Following this proposal, $\boldsymbol{w}$ results from solving the generalized eigenvalue problem $\boldsymbol{S}_b\boldsymbol{w} = \alpha\boldsymbol{S}_w\boldsymbol{w}$. Once $\boldsymbol{w}$ has been found, binary classification simply requires selecting a suitable threshold.

A well known but underexploited fact that Fisher himself pointed out [3] is that binary LDA is equivalent to the least mean squares (LMS) fitting of a hyperplane that separates ω_p and ω_n. The projection direction corresponds to the normal vector of the plane. This paper makes use of this equivalence. Aiming at image data and visual object detection, we introduce an iterative LMS approach to *separable* LDA. The resulting binary classifiers are especially suited for appearance based object recognition, because classifying image content is reducible to a convolution operation. Our practical experience has revealed several favorable characteristics of this approach. First, it is as fast as the popular cascaded weak classifiers [4]. Second, on standard databases of images of objects in complex natural scenes, it performs as reliably as recent, more sophisticated non-linear approaches [5,6,7]. Third, in contrast to the cited methods, the training time of our approach is sufficiently short to enable online learning.

W. Kropatsch, R. Sablatnig, and A. Hanbury (Eds.): DAGM 2005, LNCS 3663, pp. 318–325, 2005.

Next, we derive our algorithm and discuss its characteristics. Section 3 presents experiments on using separable LDA to detect and track objects in natural environments. A summary and an outlook will end this contribution.

2 Separable LDA for Classifying Image Data

Faced with the problem of fast and adaptive classification of image data, the idea of separable LDA arose from the following two observations.

Most approaches to appearance based object recognition transform image patches $\boldsymbol{X}$ of size $m \times n$ into vectors $\boldsymbol{x} \in \mathbb{R}^{mn}$. The first step towards fast linear discriminant analysis for visual processing is to keep the matrix representation and to consider the Frobenius inner product of matrices $\boldsymbol{W} \cdot \boldsymbol{X} = \sum_{i,j} W_{ij} X_{ij}$ instead of the inner product of high dimensional vectors.[1]

As a consequence, LDA classification of image content becomes a problem of linear filtering. If $\boldsymbol{W}$ denotes a $m \times n$ filter matrix, its convolution with a digital image $\boldsymbol{I}$ will result in a filter response map $\boldsymbol{Y}$, where an entry Y_{ij} corresponds to the LDA projection of the image patch $\boldsymbol{X}_{ij}$ centered at image coordinate (i, j), i.e. $Y_{ij} = \boldsymbol{W} \cdot \boldsymbol{X}_{ij}$.

The second step towards fast linear discriminant classification considers well known facts about linear filtering. Convolving an image with an $m \times n$ matrix requires $O(mn)$ operations per pixel. Even on modern computers, this may be prohibitive if m and n are rather large. Assume, however, $\boldsymbol{W}$ was given as a basis function expansion

$$\boldsymbol{W} = \sum_{i=1}^{k} \boldsymbol{u}_i \boldsymbol{v}_i^T \tag{1}$$

where $\boldsymbol{u}_i \in \mathbb{R}^m$ and $\boldsymbol{v}_i \in \mathbb{R}^n$ such that the basis functions are separable matrices of rank 1. Then, the two-dimensional convolution can be computed as a sequence of one-dimensional convolutions $\sum_i \big(\boldsymbol{I} * \boldsymbol{u}_i\big) * \boldsymbol{v}_i^T$. If the matrix $\boldsymbol{W}$ was rank deficient, i.e. $k < \min\{m, n\}$, this would reduce the effort to $O(k(m + n))$ and therefore would provide a fast linear approach to object detection. The following subsection discusses how to derive such separable *filter* or *projection matrices* from training data.

2.1 Learning Separable k-Term Projection Matrices

For convenience, we shall first consider the derivation of a $k = 1$ term separable LDA projection, i.e. we will examine the case $\boldsymbol{W} = \boldsymbol{u}\boldsymbol{v}^T$.

Assume a sample $\{\boldsymbol{X}_\alpha, y_\alpha\}_{\alpha=1,\ldots,N}$ of image patches $\boldsymbol{X}_\alpha$ with corresponding class labels y_α. Due to the general equivalence

$$\boldsymbol{u}\boldsymbol{v}^T \cdot \boldsymbol{X} = \sum_{k,l} (\boldsymbol{u}\boldsymbol{v}^T)_{kl} X_{kl} = \sum_{k,l} u_k v_l X_{kl} = \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{v} \tag{2}$$

a one term separable LDA projection can be found from minimizing the LMS error $E(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{2} \sum_\alpha \big(y_\alpha - \boldsymbol{u}^T \boldsymbol{X}_\alpha \boldsymbol{v}\big)^2$ using the following iterative procedure:

[1] Of course, the difference between both views is a mere conceptual one; with the common substitution $k = i \cdot n + j$ we have the equivalence $\sum_k w_k x_k = \sum_{i,j} W_{ij} X_{ij}$.

1. Randomly initialize $\boldsymbol{u} \in \mathbb{R}^m$.
2. Given $\boldsymbol{u}$, solve $E(\boldsymbol{u}, \boldsymbol{v}) = \min$ for $\boldsymbol{v}$. A solution can be found by requiring $\nabla_{\boldsymbol{v}} E(\boldsymbol{u}, \boldsymbol{v}) = 0$. If the $n \times n$ correlation matrix $\boldsymbol{C}_u = \sum_\alpha \boldsymbol{X}_\alpha^T \boldsymbol{u}\boldsymbol{u}^T \boldsymbol{X}_\alpha$ is non singular, the optimal $\boldsymbol{v}^*$ amounts to

$$\boldsymbol{v}^* = \boldsymbol{C}_u^{-1} \boldsymbol{D}_u \boldsymbol{u} \tag{3}$$

   where $\boldsymbol{D}_u = \sum_\alpha y_\alpha \boldsymbol{X}_\alpha$.
3. Given $\boldsymbol{v}^*$, solve $E(\boldsymbol{u}, \boldsymbol{v}^*) = \min$ for $\boldsymbol{u}$ in a similar way.

As the procedure starts with an arbitrary $\boldsymbol{u}$, steps 2. and 3. have to be iterated until a convergence criterion is met. Inserting (3) into the LMS error function reveals that the length of $\boldsymbol{u}$ 'cancels out'. The vector $\boldsymbol{u}$ can therefore be constrained to be of unit length $\|\boldsymbol{u}\| = 1$. Although this requires normalizing $\boldsymbol{u}$ after each iteration, it guarantees that the procedure will converge because $E(\boldsymbol{u}, \boldsymbol{v}^*)$ becomes a convex function over the unit ball in $\mathbb{R}^m$. Moreover, the unit length constraint provides a convenient convergence criterion. Our implementation uses $\|\boldsymbol{u}_t - \boldsymbol{u}_{t-1}\| \leq \epsilon$ which converges quickly.

As the resulting projection matrix $\boldsymbol{u}\boldsymbol{v}^T$ has only $m + n$ independent parameters whereas a non-separable one would provide $m \cdot n$ parameters, the one-term separable projection will be less flexible than the usual solution. This suggests we consider k-term basis expansions where $k > 1$. Note that $\boldsymbol{u}\boldsymbol{v}^T$ is of rank 1. If one demands $\boldsymbol{W} = \sum_{i=1}^k \boldsymbol{u}_i \boldsymbol{v}_i^T$ to provide more independent parameters than $\boldsymbol{u}\boldsymbol{v}^T$, it has to be of higher rank. A simple way to guarantee a higher rank of $\boldsymbol{W}$, say k, that simultaneously ensures separability of the individual terms in the basis function expansion is to require the $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ to be pairwise orthogonal, i.e $\boldsymbol{u}_i^T \boldsymbol{u}_j = 0$ and $\boldsymbol{v}_i^T \boldsymbol{v}_j = 0$ for $i \neq j$.

Forward additive stage-wise modeling [2] provides a straightforward approach to determining such sets of orthogonal parameter vectors. If $\boldsymbol{W} = \sum_{i=1}^k \boldsymbol{u}_i \boldsymbol{v}_i^T$ is a k term solution for the LDA projection matrix, a $k + 1$ term representation can be found by minimizing $E(\boldsymbol{u}_{k+1}, \boldsymbol{v}_{k+1})$. To assure orthogonality, our iterative minimization procedure has to be extended such that, after *each* iteration, the vectors $\boldsymbol{v}_{k+1}$ and $\boldsymbol{u}_{k+1}$ are orthogonalized with respect to the $\{\boldsymbol{v}_i\}_{i=1,\ldots,k}$ and $\{\boldsymbol{u}_i\}_{i=1,\ldots,k}$ determined so far. Orthogonalization can be done applying the Gram-Schmidt procedure.

Before presenting our results obtained with this approach, we first emphasize some of its favorable properties.

2.2 Properties and Benefits of Separable LDA

Separable LDA should not be confused with orthogonal LDA. Our approach does not seek a set of orthogonal discriminant directions as in the case of O-LDA [1]. Rather, we determine a discriminant direction under the constraint that the projection matrix is given as a sum of k pairwise orthogonal matrices of rank 1.

Separable LDA differs from the singular value decomposition of an unconstrained projection matrix. LMS optimization can learn an unconstrained $m \times n$ matrix $\boldsymbol{W}$. Using SVD, it can be decomposed into a sum of r separable rank 1 matrices, where r is the rank of $\boldsymbol{W}$. Since $\boldsymbol{W}$ is usually of full rank, $r = \min\{m, n\}$. If, w.l.o.g., we assume $r = m$, a separated convolution will require $m(m + n) > mn$ operations per

pixel and will thus be even more expensive than the usual variant. A rank deficient SVD expansion of $k < r$ terms is less expensive, but practical experience shows that it not useful. In contrast to the SVD-based method, our approach derives the projection matrix directly from data rather than from an unconstrained $\boldsymbol{W}$. Hence, even for $k \ll r$, it yields reasonable results and also runs quickly.

Separable LDA differs from 2D LDA, as introduced by Ye et al. [8], who present an iterative SVD algorithm that projects $m \times n$ matrices onto $l_1 \times l_2$ matrices where $l_1 < m$ and $l_2 < n$. For projections onto a one-dimensional subspace ($l_1, l_2 = 1$), their approach is equivalent to our solution for 1-term separable LDA. However, as it does not allow for $k > 1$ term representations, their algorithm provides fever independent parameters than our approach to binary LDA.

Separable LDA differs from image coding using the tensor rank principle proposed by Shashua and Levin [9]. Although it resembles our approach, their algorithm has a fundamentally different purpose and considers a different optimization criterion. While separable LDA seeks a k-term projection matrix of low rank, Shashua and Levin estimate a minimal set of second order tensors of rank 1 having a linear span that includes the given set of training images.

Separable LDA projection matrices are learned quickly. LDA based on unconstrained LMS optimization or on solving the generalized eigenvalue problem $\boldsymbol{S}_b\boldsymbol{w} = \alpha\boldsymbol{S}_w\boldsymbol{w}$ requires the computation and inversion of covariance matrices of sizes $mn \times mn$. For larger values of m and n and many training examples, training becomes tedious, even on modern computers. However, the covariance matrices $\boldsymbol{C}_u$ and $\boldsymbol{C}_v$ that appear in the learning stage of separable LDA are of considerably reduced sizes $n \times n$ and $m \times m$, respectively. Therefore, in addition to its fast runtime, our technique significantly shortens training time.

Separable LDA tackles the small sample size problem. This property is closely related to the previous one. The term *small sample size problem* refers to the effect that the within-class scatter matrix $\boldsymbol{S}_w$ is often singular because the number of training samples is much smaller than the dimension of the embedding space [1]. Again, as the covariance matrices $\boldsymbol{C}_u$ and $\boldsymbol{C}_v$ are of considerably small dimensionality, small sample sizes will not hamper separable LDA.

Separable LDA can be expected to perform well in visual object detection and recognition. Fast training and operation times allow the use of fairly large values for m and n so that the resulting linear classifiers will process data from very high dimensional feature spaces. However, according to Cover's theorem [10], the probability of finding a suitable hyperplane that separates two arbitrary classes increases with the dimension of the embedding space.

3 Experiments

This section presents two application examples for the algorithm derived above. First, we regard the problem of robust offline object detection in real world environments. Afterwards, we consider online learning for tracking of articulated objects.

Note that in all experiments the input was normalized to zero mean $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{\mu}$, where $\boldsymbol{\mu}$ denotes the mean of all training examples. In the classification stage, this

Fig. 1. Exemplary detection results (k = 9, θ = 3.6) obtained on the UIUC database of cars [5]

accounts only for a single operation per pixel, since $(\boldsymbol{X} - \boldsymbol{\mu}) \cdot \boldsymbol{W} = \boldsymbol{X} \cdot \boldsymbol{W} - \boldsymbol{\mu} \cdot \boldsymbol{W}$, where the constant $\boldsymbol{\mu} \cdot \boldsymbol{W}$ can be computed beforehand.

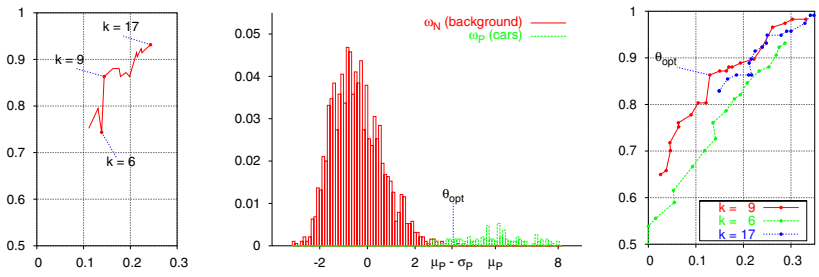
With respect to the classification threshold θ applied in our experiments, we must point out that determining the theoretically optimal threshold for LDA-based classification requires knowledge of the class covariance matrices. However, in the previous section, we saw that separable LDA avoids the estimation of class covariance matrices, and we stressed the advantages this entails. In order not to lose these advantages, we adopted a heuristic to automatically determine a suitable θ. We computed the mean μ_p and variance σ_p from the projections of the positive training samples onto the discriminant direction, and θ was set to $\mu_p - \sigma_p$. Figure 2(b) indicates that this is good practice in many cases.

3.1 Object Detection in Real World Environments

Experiments in offline object detection were carried out using the UIUC database of cars [5,11]. It contains side views of cars of arbitrary shape and color in natural environments; typical examples are shown in Figure 1. In our experiments, the set ω_p of positive training examples (label +1) consisted of 124 images of cars of size 80×30. The set ω_n of negative examples (label -1) consisted of 1776 patches randomly cut from the background of half of the images in the database; testing was done on the other half.

In order to reduce effects of varying illumination and color, the classifiers considered in our experiments were trained and applied to gradient magnitude images. These were obtained using recursive Gaussian filtering according to Deriche [12]. After training, the test images were convolved with the resulting matrices. A car was said to be detected where the classifier response exceeded the threshold θ. The response maps were subjected to a non-maximum suppression to reduce the number of false positives.

As the UIUC database includes manually annotated ground truth, detection results can be colored correspondingly (see Figure 1). Figure 2(a) plots the precision recall curve we obtained for classifiers of different rank k. The $k = 9$ classifier yielded the best ratio of recall and precision. Figure 2(b) shows how it discriminates the training images. For the classifiers $k = 9$, $k = 6$ (lowest recall) and $k = 17$ (highest recall) we varied the classification thresholds in the intervals $[\theta(k) - 1, \theta(k) + 1]$ to examine how it would influence the performance. The resulting precision recall curves are shown in Figure 2(c). It turned out that improvements were possible. In terms of equal error rate (EER), i.e. the point of equal recall and precision, the classifier with $k = 9$ and $\theta = 3.6$ performed best; it yielded an EER of 86%.



(a) (b) (c)

Fig. 2. Quantitative results obtained on the UIUC database of cars. 2(a) $1-$precision vs. recall for different k; the classification threshold θ was set to $\mu_p(k) - \sigma_p(k)$. 2(b) histogram resulting from projecting the training data using $k = 9$. 2(c) $1-$precision vs. recall curves obtained from varying θ for $k \in \{6, 9, 17\}$; the classifier with $k = 9$ and $\theta_{\text{opt}} = 3.6$ provides an EER of 86%.

Table 1. Comparison of results reported in contributions dealing with the UIUC database of cars

method	Agarwal et al.[5]	**separable LDA**	Fergus et al. [6]	Garg et al. [7]	Leibe et al. [13]
EER	77%	86%	88%	88%	97%

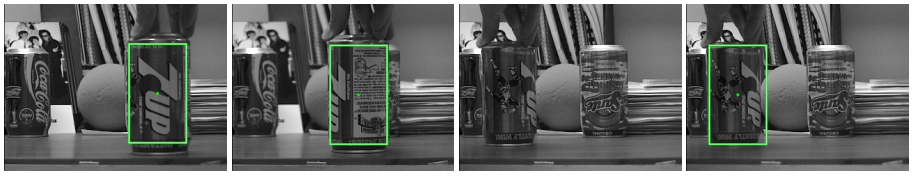


Fig. 3. Exemplary results ($k = 7$, $\lambda = 10$, $\theta = \mu_p - \sigma_p$) obtained on frames 41, 89, 304 and 310 of the rotating can sequence recorded by Black and Jepson [14].

It is interesting to note that our linear and holistic approach performs comparably to sophisticated methods found in recent literature. Table 1 lists equal error rates other researchers reported for the UIUC database. Except for our method, all figures result from part-based approaches that learn lexica of salient object parts and statistical models of part relations. Runtimes or training times of these approaches have not been reported but, due to the need for building lexica, at least the training times can be expected to exceed real time. Given a naïve C implementation, our method performed as follows: on a 3GHz Xeon PC, training with 1900 examples took 13 seconds. File I/O and processing of 99 test images of an average size of 120×116 pixels was done at a rate of 3.2Hz.

3.2 Online Learning for Tracking of Articulated Objects

Encouraged by the runtime behavior of our approach, we explored its potential in online learning from image sequences and experimented with material provided by Black and Jepson [14]. Next, we discuss a typical example.

The *rotating can* sequence shows a tin can that is being moved and rotated in front of a static camera (see Figure 3). For our experiments, we applied 141×81 image patches to train classifiers of various ranks k. After manually specifying the center of the can in the first frame of the sequence, 30 image patches are randomly selected from its neighborhood to serve as positive training examples (class ω_p, label +1); 240 patches randomly selected from outside the neighborhood were used as counter examples (class ω_n, label -1). Training and classification both considered simple pixel intensity information.

The classifier matrix resulting from training on the first frame was convolved with the subsequent frames of the sequence. This was done in a brute force manner: the entire image was processed, without considering regions of interest. The can was assumed to be recognized where the classifier response exceeded the threshold θ. Again, non-maximum suppression was applied to reduce the number of false positives. Due to its rotation, the can's appearance changes throughout the sequence. Thus, after λ frames, each classifier was retrained. We experimented with $\lambda \in \{3, 6, 9, \ldots, 30\}$.

The graphs in Figure 4 show 1-precision and recall curves for classifiers of rank $k \in \{4, 7, 10\}$. They are plotted as functions of the operation frequency, which, in turn, is a function of λ. As one would expect, the 4 term classifiers perform fastest. Owing to the needs of video processing, we improved the memory management of our implementation. Consequently, on a 3GHz Xeon PC, the 400 frames of size 320×240 were processed at an operation frequency of up to 9Hz, including file I/O *and* retraining. A practically suitable ratio of speed and reliability was obtained for a $k = 7$ classifier retrained every 9 frames. At a frequency of 4.3Hz, it produces a recall of 94% and a 1-precision value of 3%.

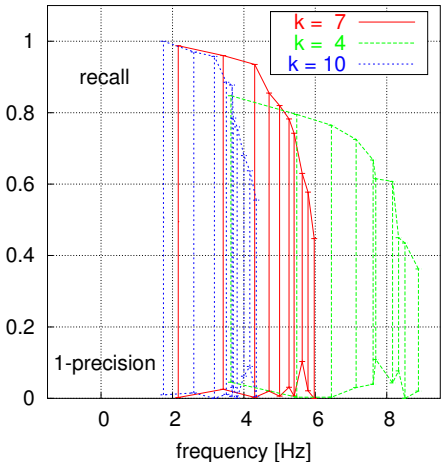


Fig. 4. Quantitative results obtained on the rotating can sequence. The graphs show recall and $1-$precision for classifiers of different k plotted as functions of the operation frequencies of various update rates λ. The k = 7 classifier updated every 9 frames operates at 4.3 Hz and produces a recall of 94% and $1-$precision of 3%.

4 Summary and Outlook

This paper presented an approach to separable linear discriminant classification for image analysis. Based on the idea of understanding LDA projection as a convolution operation, we express the projection matrix as a basis function expansion of separable rank 1 matrices, in order to ensure rapid runtime behavior. We introduced an iterative two step least mean squares procedure to learn corresponding basis functions from training data. Due to the separability of the projection operator, both application

and training are very fast. Furthermore, small sample sizes do not corrupt the training process. Experimental results obtained on a standard testbed for object detection revealed that separable LDA performs as fast and as reliably as more elaborate state-of-the-art techniques. In addition, however, it also provides an avenue to online learning in image sequence processing.

Currently, we are working on a thorough experimental and theoretical analysis of separable LDA. Our focus is on questions to which this paper only alludes: Is there a significant difference in performance between usual binary LDA and k-term separable LDA? How can a suitable number k of terms be determined automatically How can our approach be extended to multiple classes? Is there a framework that could unify our approach to computing k-term separable matrices with the approaches proposed in [8] and [9] that consider 1-term higher order tensors of rank 1?

Acknowledgments

We would like to thank Andrei Rotenstein for fruitful discussions and valuable suggestions. We also want to thank our anonymous reviewers. Their precise comments and pointers to recent related literature helped to improve this contribution.

References

1. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press (1990)
2. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer (2001)
3. Fisher, R.: The Use of Multiple Measurements in Taxonomic Problems. Ann. Eugenics **7** (1936) 179–188
4. Viola, P., Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: Proc. CVPR. Volume I. (2001) 511–518
5. Agarwal, S., Awan, A., Roth, D.: Learning to Detect Objects in Images via a Sparse, Part-based Representation. IEEE T. Pattern Anal. Machine Intell. **26** (2004) 1475–1490
6. Fergus, R., Perona, P., Zisserman, A.: Object Class Recognition by Unsupervised Scale-Invariant Learning. In: Proc. CVPR. Volume II. (2003) 264–272
7. Garg, A., .Agarwal, S., Huang, T.: Fusion of Global and Local Information for Object Detection. In: Proc. ICPR. Volume III. (2002) 723–727
8. Ye, J., Janardan, R., Li, Q.: Two-Dimensional Linear Discriminant Analysis. In Saul, L., Weiss, Y., Bottou, L., eds.: Advances in Neural Information Processing Systems 17. MIT Press, Cambridge, MA (2005) 1569–1576
9. Shashua, A., Levin, A.: Linear Image Cosing for Regression and Classification using the Tensor-rank Principle. In: Proc. CVPR. Volume I. (2001) 42–40
10. Cover, T.: Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications to Pattern Recognition. IEEE T. on Electronic Computers **14** (1965) 326–334
11. http://l2r.cs.uiuc.edu/˜cogcomp/Data/Car/ (retreived spring 2005)
12. Deriche, R.: Recursively Implementing the Gaussian and Its Derivatives. In: Proc. ICIP. (1992) 263–267
13. Leibe, B., Schiele, B.: Scale-Invariant Object Categorization using a Scale-Adaptive Mean-Shift Search. In: Proc. DAGM. Volume 3175 of LNCS., Springer (2004) 145–153
14. Black, M., Jepson, A.: EigenTracking: Robust matching and tracking of articulated objects using a view-based reprenstation. Int. J. Comput. Vis. **26** (1998) 63–84