Joab Winkler
Mahesan Niranjan
Neil Lawrence (Eds.)

LNAI 3635

# Deterministic and Statistical Methods in Machine Learning

First International Workshop
Sheffield, UK, September 2004
Revised Lectures

Springer

Joab Winkler Mahesan Niranjan
Neil Lawrence (Eds.)

# Deterministic and Statistical Methods in Machine Learning

First International Workshop
Sheffield, UK, September 7-10, 2004
Revised Lectures

Springer

Volume Editors

Joab Winkler
Mahesan Niranjan
Neil Lawrence
The University of Sheffield, Department of Computer Science
Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK
E-mail: Winkler@dcs.shef.ac.uk
M.Niranjan@sheffield.ac.uk
neil@dcs.shef.ac.uk

Library of Congress Control Number: 2005933155

CR Subject Classification (1998): I.2, F.2.2, I.5, I.4, F.4.1, H.3

ISSN 0302-9743
ISBN-10 3-540-29073-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-29073-5 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11559887 06/3142 5 4 3 2 1 0

Lecture Notes in Artificial Intelligence 3635

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

# Lecture Notes in Artificial Intelligence (LNAI)

Vol. 3735: A. Hoffmann, H. Motoda, T. Scheffer (Eds.), Discovery Science. XVI, 400 pages. 2005.

Vol. 3734: S. Jain, H.U. Simon, E. Tomita (Eds.), Algorithmic Learning Theory. XII, 490 pages. 2005.

Vol. 3721: A. Jorge, L. Torgo, P. Brazdil, R. Camacho, J. Gama (Eds.), Knowledge Discovery in Databases: PKDD 2005. XXIII, 719 pages. 2005.

Vol. 3720: J. Gama, R. Camacho, P. Brazdil, A. Jorge, L. Torgo (Eds.), Machine Learning: ECML 2005. XXIII, 769 pages. 2005.

Vol. 3717: B. Gramlich (Ed.), Frontiers of Combining Systems. X, 321 pages. 2005.

Vol. 3702: B. Beckert (Ed.), Automated Reasoning with Analytic Tableaux and Related Methods. XIII, 343 pages. 2005.

Vol. 3698: U. Furbach (Ed.), KI 2005: Advances in Artificial Intelligence. XIII, 409 pages. 2005.

Vol. 3690: M. Pěchouček, P. Petta, L.Z. Varga (Eds.), Multi-Agent Systems and Applications IV. XVII, 667 pages. 2005.

Vol. 3684: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part IV. LXXIX, 933 pages. 2005.

Vol. 3683: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part III. LXXX, 1397 pages. 2005.

Vol. 3682: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part II. LXXIX, 1371 pages. 2005.

Vol. 3681: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part I. LXXX, 1319 pages. 2005.

Vol. 3673: S. Bandini, S. Manzoni (Eds.), AI*IA 2005: Advances in Artificial Intelligence. XIV, 614 pages. 2005.

Vol. 3662: C. Baral, G. Greco, N. Leone, G. Terracina (Eds.), Logic Programming and Nonmonotonic Reasoning. XIII, 454 pages. 2005.

Vol. 3661: T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, T. Rist (Eds.), Intelligent Virtual Agents. XIII, 506 pages. 2005.

Vol. 3658: V. Matoušek, P. Mautner, T. Pavelka (Eds.), Text, Speech and Dialogue. XV, 460 pages. 2005.

Vol. 3651: R. Dale, K.-F. Wong, J. Su, O.Y. Kwong (Eds.), Natural Language Processing – IJCNLP 2005. XXI, 1031 pages. 2005.

Vol. 3642: D. Ślezak, J. Yao, J.F. Peters, W. Ziarko, X. Hu (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Part II. XXIII, 738 pages. 2005.

Vol. 3641: D. Ślezak, G. Wang, M. Szczuka, I. Düntsch, Y. Yao (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Part I. XXIV, 742 pages. 2005.

Vol. 3635: J. Winkler, M. Niranjan, N. Lawrence (Eds.), Deterministic and Statistical Methods in Machine Learning. VIII, 341 pages. 2005.

Vol. 3632: R. Nieuwenhuis (Ed.), Automated Deduction – CADE-20. XIII, 459 pages. 2005.

Vol. 3630: M.S. Capcarrere, A.A. Freitas, P.J. Bentley, C.G. Johnson, J. Timmis (Eds.), Advances in Artificial Life. XIX, 949 pages. 2005.

Vol. 3626: B. Ganter, G. Stumme, R. Wille (Eds.), Formal Concept Analysis. X, 349 pages. 2005.

Vol. 3625: S. Kramer, B. Pfahringer (Eds.), Inductive Logic Programming. XIII, 427 pages. 2005.

Vol. 3620: H. Muñoz-Avila, F. Ricci (Eds.), Case-Based Reasoning Research and Development. XV, 654 pages. 2005.

Vol. 3614: L. Wang, Y. Jin (Eds.), Fuzzy Systems and Knowledge Discovery, Part II. XLI, 1314 pages. 2005.

Vol. 3613: L. Wang, Y. Jin (Eds.), Fuzzy Systems and Knowledge Discovery, Part I. XLI, 1334 pages. 2005.

Vol. 3607: J.-D. Zucker, L. Saitta (Eds.), Abstraction, Reformulation and Approximation. XII, 376 pages. 2005.

Vol. 3596: F. Dau, M.-L. Mugnier, G. Stumme (Eds.), Conceptual Structures: Common Semantics for Sharing Knowledge. XI, 467 pages. 2005.

Vol. 3593: V. Mařík, R. W. Brennan, M. Pěchouček (Eds.), Holonic and Multi-Agent Systems for Manufacturing. XI, 269 pages. 2005.

Vol. 3587: P. Perner, A. Imiya (Eds.), Machine Learning and Data Mining in Pattern Recognition. XVII, 695 pages. 2005.

Vol. 3584: X. Li, S. Wang, Z.Y. Dong (Eds.), Advanced Data Mining and Applications. XIX, 835 pages. 2005.

Vol. 3581: S. Miksch, J. Hunter, E. Keravnou (Eds.), Artificial Intelligence in Medicine. XVII, 547 pages. 2005.

Vol. 3577: R. Falcone, S. Barber, J. Sabater-Mir, M.P. Singh (Eds.), Trusting Agents for Trusting Electronic Societies. VIII, 235 pages. 2005.

Vol. 3575: S. Wermter, G. Palm, M. Elshaw (Eds.), Biomimetic Neural Learning for Intelligent Robots. IX, 383 pages. 2005.

Vol. 3571: L. Godo (Ed.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty. XVI, 1028 pages. 2005.

Vol. 3559: P. Auer, R. Meir (Eds.), Learning Theory. XI, 692 pages. 2005.

Vol. 3558: V. Torra, Y. Narukawa, S. Miyamoto (Eds.), Modeling Decisions for Artificial Intelligence. XII, 470 pages. 2005.

Vol. 3554: A. Dey, B. Kokinov, D. Leake, R. Turner (Eds.), Modeling and Using Context. XIV, 572 pages. 2005.

Vol. 3550: T. Eymann, F. Klügl, W. Lamersdorf, M. Klusch, M.N. Huhns (Eds.), Multiagent System Technologies. XI, 246 pages. 2005.

Vol. 3539: K. Morik, J.-F. Boulicaut, A. Siebes (Eds.), Local Pattern Detection. XI, 233 pages. 2005.

Vol. 3538: L. Ardissono, P. Brna, A. Mitrovic (Eds.), User Modeling 2005. XVI, 533 pages. 2005.

Vol. 3533: M. Ali, F. Esposito (Eds.), Innovations in Applied Artificial Intelligence. XX, 858 pages. 2005.

Vol. 3528: P.S. Szczepaniak, J. Kacprzyk, A. Niewiadomski (Eds.), Advances in Web Intelligence. XVII, 513 pages. 2005.

Vol. 3518: T.B. Ho, D. Cheung, H. Liu (Eds.), Advances in Knowledge Discovery and Data Mining. XXI, 864 pages. 2005.

Vol. 3508: P. Bresciani, P. Giorgini, B. Henderson-Sellers, G. Low, M. Winikoff (Eds.), Agent-Oriented Information Systems II. X, 227 pages. 2005.

Vol. 3505: V. Gorodetsky, J. Liu, V. Skormin (Eds.), Autonomous Intelligent Systems: Agents and Data Mining. XIII, 303 pages. 2005.

Vol. 3501: B. Kégl, G. Lapalme (Eds.), Advances in Artificial Intelligence. XV, 458 pages. 2005.

Vol. 3492: P. Blache, E. Stabler, J. Busquets, R. Moot (Eds.), Logical Aspects of Computational Linguistics. X, 363 pages. 2005.

Vol. 3490: L. Bolc, Z. Michalewicz, T. Nishida (Eds.), Intelligent Media Technology for Communicative Intelligence. X, 259 pages. 2005.

Vol. 3488: M.-S. Hacid, N.V. Murray, Z.W. Raś, S. Tsumoto (Eds.), Foundations of Intelligent Systems. XIII, 700 pages. 2005.

Vol. 3487: J. Leite, P. Torroni (Eds.), Computational Logic in Multi-Agent Systems. XII, 281 pages. 2005.

Vol. 3476: J. Leite, A. Omicini, P. Torroni, P. Yolum (Eds.), Declarative Agent Languages and Technologies II. XII, 289 pages. 2005.

Vol. 3464: S.A. Brueckner, G.D.M. Serugendo, A. Karageorgos, R. Nagpal (Eds.), Engineering Self-Organising Systems. XIII, 299 pages. 2005.

Vol. 3452: F. Baader, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XI, 562 pages. 2005.

Vol. 3451: M.-P. Gleizes, A. Omicini, F. Zambonelli (Eds.), Engineering Societies in the Agents World V. XIII, 349 pages. 2005.

Vol. 3446: T. Ishida, L. Gasser, H. Nakashima (Eds.), Massively Multi-Agent Systems I. XI, 349 pages. 2005.

Vol. 3445: G. Chollet, A. Esposito, M. Faundez-Zanuy, M. Marinaro (Eds.), Nonlinear Speech Modeling and Applications. XIII, 433 pages. 2005.

Vol. 3438: H. Christiansen, P.R. Skadhauge, J. Villadsen (Eds.), Constraint Solving and Language Processing. VIII, 205 pages. 2005.

Vol. 3430: S. Tsumoto, T. Yamaguchi, M. Numao, H. Motoda (Eds.), Active Mining. XII, 349 pages. 2005.

Vol. 3419: B. Faltings, A. Petcu, F. Fages, F. Rossi (Eds.), Recent Advances in Constraints. X, 217 pages. 2005.

Vol. 3416: M. Böhlen, J. Gamper, W. Polasek, M.A. Wimmer (Eds.), E-Government: Towards Electronic Democracy. XIII, 311 pages. 2005.

Vol. 3415: P. Davidsson, B. Logan, K. Takadama (Eds.), Multi-Agent and Multi-Agent-Based Simulation. X, 265 pages. 2005.

Vol. 3403: B. Ganter, R. Godin (Eds.), Formal Concept Analysis. XI, 419 pages. 2005.

Vol. 3398: D.-K. Baik (Ed.), Systems Modeling and Simulation: Theory and Applications. XIV, 733 pages. 2005.

Vol. 3397: T.G. Kim (Ed.), Artificial Intelligence and Simulation. XV, 711 pages. 2005.

Vol. 3396: R.M. van Eijk, M.-P. Huget, F. Dignum (Eds.), Agent Communication. X, 261 pages. 2005.

Vol. 3394: D. Kudenko, D. Kazakov, E. Alonso (Eds.), Adaptive Agents and Multi-Agent Systems II. VIII, 313 pages. 2005.

Vol. 3392: D. Seipel, M. Hanus, U. Geske, O. Bartenstein (Eds.), Applications of Declarative Programming and Knowledge Management. X, 309 pages. 2005.

Vol. 3374: D. Weyns, H. V.D. Parunak, F. Michel (Eds.), Environments for Multi-Agent Systems. X, 279 pages. 2005.

Vol. 3371: M.W. Barley, N. Kasabov (Eds.), Intelligent Agents and Multi-Agent Systems. X, 329 pages. 2005.

Vol. 3369: V. R. Benjamins, P. Casanovas, J. Breuker, A. Gangemi (Eds.), Law and the Semantic Web. XII, 249 pages. 2005.

Vol. 3366: I. Rahwan, P. Moraitis, C. Reed (Eds.), Argumentation in Multi-Agent Systems. XII, 263 pages. 2005.

Vol. 3359: G. Grieser, Y. Tanaka (Eds.), Intuitive Human Interfaces for Organizing and Accessing Intellectual Assets. XIV, 257 pages. 2005.

Vol. 3346: R.H. Bordini, M. Dastani, J. Dix, A.E.F. Seghrouchni (Eds.), Programming Multi-Agent Systems. XIV, 249 pages. 2005.

Vol. 3345: Y. Cai (Ed.), Ambient Intelligence for Scientific Discovery. XII, 311 pages. 2005.

Vol. 3343: C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, T. Barkowsky (Eds.), Spatial Cognition IV. XIII, 519 pages. 2005.

Vol. 3339: G.I. Webb, X. Yu (Eds.), AI 2004: Advances in Artificial Intelligence. XXII, 1272 pages. 2004.

Vol. 3336: D. Karagiannis, U. Reimer (Eds.), Practical Aspects of Knowledge Management. X, 523 pages. 2004.

Vol. 3327: Y. Shi, W. Xu, Z. Chen (Eds.), Data Mining and Knowledge Management. XIII, 263 pages. 2005.

Vol. 3315: C. Lemaître, C.A. Reyes, J.A. González (Eds.), Advances in Artificial Intelligence – IBERAMIA 2004. XX, 987 pages. 2004.

Vol. 3303: J.A. López, E. Benfenati, W. Dubitzky (Eds.), Knowledge Exploration in Life Science Informatics. X, 249 pages. 2004.

¥453.12元

# Preface

Machine learning is a rapidly maturing field that aims to provide practical methods for data discovery, categorization and modelling. The Sheffield Machine Learning Workshop, which was held 7–10 September 2004, brought together some of the leading international researchers in the field for a series of talks and posters that represented new developments in machine learning and numerical methods.

The workshop was sponsored by the Engineering and Physical Sciences Research Council (EPSRC) and the London Mathematical Society (LMS) through the MathFIT program, whose aim is the encouragement of new interdisciplinary research. Additional funding was provided by the PASCAL European Framework 6 Network of Excellence and the University of Sheffield. It was the commitment of these funding bodies that enabled the workshop to have a strong program of invited speakers, and the organizers wish to thank these funding bodies for their financial support. The particular focus for interactions at the workshop was *Advanced Research Methods in Machine Learning and Statistical Signal Processing.*

These proceedings contain work that was presented at the workshop, and ideas that were developed through, or inspired by, attendance at the workshop. The proceedings reflect this mixture and illustrate the diversity of applications and theoretical work in machine learning.

We would like to thank the presenters and attendees at the workshop for the excellent quality of presentation and discussion during the oral and poster sessions. We are also grateful to Gillian Callaghan for her support in the organization of the workshop, and finally we wish to thank the anonymous reviewers for their help in compiling the proceedings.

July 2005

Joab Winkler<br>
Neil Lawrence<br>
Mahesan Niranjan

# Table of Contents

# Object Recognition via Local Patch Labelling

Christopher M. Bishop[1] and Ilkay Ulusoy[2]

[1] Microsoft Research,
7 J J Thompson Avenue,
Cambridge, UK
http://research.microsoft.com/~cmbishop
[2] METU, Computer Vision and Intelligent Systems Research Lab.,
06531 Ankara, Turkey
http://www.eee.metu.edu.tr/~ilkay

**Abstract.** In recent years the problem of object recognition has received considerable attention from both the machine learning and computer vision communities. The key challenge of this problem is to be able to recognize any member of a category of objects in spite of wide variations in visual appearance due to variations in the form and colour of the object, occlusions, geometrical transformations (such as scaling and rotation), changes in illumination, and potentially non-rigid deformations of the object itself. In this paper we focus on the detection of objects within images by combining information from a large number of small regions, or 'patches', of the image. Since detailed hand-segmentation and labelling of images is very labour intensive, we make use of 'weakly labelled' data in which the training images are labelled only according to the presence or absence of each category of object. A major challenge presented by this problem is that the foreground object is accompanied by widely varying background clutter, and the system must learn to distinguish the foreground from the background without the aid of labelled data. In this paper we first show that patches which are highly relevant for the object discrimination problem can be selected automatically from a large dictionary of candidate patches during learning, and that this leads to improved classification compared to direct use of the full dictionary. We then explore alternative techniques which are able to provide labels for the individual patches, as well as for the image as a whole, so that each patch is identified as belonging to one of the object categories or to the background class. This provides a rough indication of the location of the object or objects within the image. Again these individual patch labels must be learned on the basis only of overall image class labels. We develop two such approaches, one discriminative and one generative, and compare their performance both in terms of patch labelling and image labelling. Our results show that good classification performance can be obtained on challenging data sets using only weak training labels, and they also highlight some of the relative merits of discriminative and generative approaches.

## 1 Introduction

The problem of object recognition has emerged as a 'grand challenge' for computer vision, with the longer term aim of being able to achieve near human levels of recognition for tens of thousands of object categories under a wide variety of conditions. Many of

J. Winkler, N. Lawrence, M. Niranjan (Eds.): Machine Learning Workshop, LNAI 3635, pp. 1–21, 2005.

the current approaches to this problem rely on the use of local features obtained from small patches of the image. The motivation for this is that the variability of small patches is much less than that of whole images and so there are much better prospects for generalization, in other words for recognizing that a patch from a test image is similar to patches in the training images. However, the patches must be sufficiently variable, and therefore sufficiently large, to be able to discriminate between the different object categories and also between objects and background clutter. A good way to balance these two conflicting requirements is to determine the object categories present in an image by fusing together partial ambiguous information from multiple patches. Probability theory provides a powerful framework for combining such uncertain information in a principled manner, and will form the basis for our research (the specific local features that we use in this paper are described in Section 2.) Also, the locations of those patches which provide strong evidence for an object also give an indication of the location and spatial extent of that object.

In common with a number of previous approaches, we do not attempt to model the spatial relationship between patches. Although such spatial information is certainly very relevant to the object recognition problem, and its inclusion would be expected to improved recognition performance for many object categories, its role is complementary to that of the texture-like evidence provided by local patches. Here we show that local information alone can already give good discriminatory results.

A key issue in object recognition is the need for predictions to be invariant to a wide variety of transformations of the input image due to translations and rotations of the object in 3D space, changes in viewing direction and distance, variations in the intensity and nature of the illumination, and non-rigid transformations of the object. Although the informative features used in [13] are shown to be superior to generic features when used with a simple classification method, they are not invariant to scale and orientation. By contrast, generic interest point operators such as saliency [6], DoG [7] and Harris-Laplace [9] detectors are repeatable in the sense that they are invariant to location, scale and orientation, and some are also affine invariant [7,9] to some extent. For the purposes of this paper we shall consider the use of invariant features obtained from local regions of the image centered on interest points.

Fergus et al. [5] learn jointly the appearances and relative locations of a small set of parts whose potential locations are determined by a saliency detector [6]. Since their algorithm is very complex, the number of parts has to be kept small and the type of detector they used is appropriate for this purpose. Csurka *et al.* [3] used Harris-Laplace interest point operators [9] with SIFT features [7] for the purpose of multi class object category recognition. Features are clustered using K-Means and each feature is labelled according to the closest cluster centre. Histograms of feature labels are then used as class-conditional densities. Since such interest point operators detect many points from the background as well as from the object itself, the features are used collectively to determine the object category, and no information on object localization is obtained. In [4], informative features were selected based on information criteria such as likelihood ratio and mutual information in which DoG and Harris-Laplace interest point detectors with SIFT descriptors were compared. However, in this supervised approach, hundreds of images were hand segmented in order to train support vector machine and Gaussian mixture models (GMMs) for foreground/background classification. The two detectors gave similar results although DoG produces more features from the background. Finally, Xie and Perez [14] extended the GMM based approach of [4] to a semi-supervised case inspired from [5]. A multi-modal GMM was trained to model foreground and background features where some uncluttered images of foreground were used for the purpose of initialization.

In this paper we develop several new approaches to object recognition based on features extracted from local patches centered on interest points. We begin, in Section 3, by extending the model of [3] which constructs a large dictionary of candidate feature 'prototypes'. By using the technique of *automatic relevance determination*, our approach can learn which of these prototypes are particularly salient for the problem of discriminating object classes and can thereby give appropriately less emphasis to those which carry little discriminatory information (such as those associated with background clutter). This leads to a significant improvement in classification performance.

While this approach allows the system to focus on the foreground objects, it does not directly lead to a labelling of the individual patches. We therefore develop new probabilistic approaches to object recognition based on local patches in which the system learns not only to classify the overall image, but also to assign labels to patches themselves. In particular, we develop two complementary approaches one of which is discriminative (Section 4) and one of which is generative (Section 5).

To understand the distinction between discriminative and generative, consider a scenario in which an image described by a vector $\mathbf{X}$ (which might comprise raw pixel intensities, or some set of features extracted from the image) is to be assigned to one of $K$ classes $k = 1, \ldots, K$. From basic decision theory [2] we know that the most complete characterization of the solution is expressed in terms of the set of posterior probabilities $p(k|\mathbf{X})$. Once we know these probabilities it is straightforward to assign the image $\mathbf{X}$ to a particular class to minimize the expected loss (for instance, if we wish to minimize the number of misclassifications we assign $\mathbf{X}$ to the class having the largest posterior probability).

In a discriminative approach we introduce a parametric model for the posterior probabilities, and infer the values of the parameters from a set of labelled training data. This may be done by making point estimates of the parameters using maximum likelihood, or by computing distributions over the parameters in a Bayesian setting (for example by using variational inference).

By contrast, in a generative approach we model the joint distribution $p(k, \mathbf{X})$ of images and labels. This can be done, for instance, by learning the class prior probabilities $p(k)$ and the class-conditional densities $p(\mathbf{X}|k)$ separately. The required posterior probabilities are then obtained using Bayes' theorem

$$p(k|\mathbf{X}) = \frac{p(\mathbf{X}|k)p(k)}{\sum_j p(\mathbf{X}|j)p(j)} \tag{1}$$

where the sum in the denominator is taken over all classes.

Comparative results from the various approaches are presented in Section 6. These show that the generative approach gives excellent classification performance both for individual patches and for the complete images, but that careful initialization of the training procedure is required. By contrast the discriminative approach, which gives good results for image labelling but not for patch labelling, is significantly faster in processing test images. Ideas for future work, including techniques for combining the benefits of generative and discriminative approaches, are discussed briefly in Section 7.

## 2 Local Feature Extraction

Our goal in this paper is not to find optimal features and representations for solving a specific object recognition task, but rather to fix on a particular, widely used, feature set and use this as the basis to compare alternative learning methodologies. We shall also fix on a specific data set, chosen for the wide variability of the objects in order to present a non-trivial classification problem. In particular, we consider the task of detecting and distinguishing cows and sheep in natural images.

We therefore follow several recent approaches [7,9] and use an interest point detector to focus attention on a small number of local patches in each image. This is followed by invariant feature extraction from a neighbourhood around each interest point. Specifically we use DoG interest point detectors, and at each interest point we extract a 128 dimensional SIFT feature vector [7] from a patch whose scale is determined by the DoG detector. Following [1] we concatenate the SIFT features with additional colour features comprising average and standard deviation of $(R, G, B)$, $(L, a, b)$ and $(r = R/(R+G+B), g = G/(R+G+B))$, which gives an overall 144 dimensional feature vector. The result of applying the DoG operator to a cow image is shown in Figure 1.

In this paper we use $\mathbf{t}_n$ to denote the image label vector for image $n$ with independent components $t_{nk} \in \{0, 1\}$ in which $k = 1, \ldots K$ labels the class. Each class can be present or absent independently in an image, and we make no distinction between foreground and background classes within the model itself. $\mathbf{X}_n$ denotes the observation for image $n$ and this comprises as set of $J_n$ patch vectors $\{\mathbf{x}_{nj}\}$ where $j = 1, \ldots, J_n$. Note that the number $J_n$ of detected interest points will in general vary from image to image.

On a small-scale problem it is reasonable to segment and label the objects present in the training images. However, for large-scale object recognition involving thousands of categories this will not be feasible, and so instead it is necessary to employ training data which is at best 'weakly labelled'. Here we consider a training set in which each image is labelled only according to the presence or absence of each category of object (in our example each image contains either cows or sheep).

## 3 Patch Saliency Using Automatic Relevance Determination

We begin by considering a simple approach based on [3]. In this method the features extracted from all of the training images are clustered into $C$ classes using the K-means algorithm, after which each patch in each image is assigned to the closest prototype. Each image $n$ is therefore described by a fixed-length histogram feature vector $\mathbf{h}_n$ of length $C$ in which element $h_{nc}$ represents the number of patches in image $n$ which are assigned to cluster $c$, where $c \in \{1, \ldots, C\}$ and $n \in \{1, \ldots, N\}$. These feature vectors are then used to construct a classifier which takes an image $\mathbf{X}_n$ as input, converts

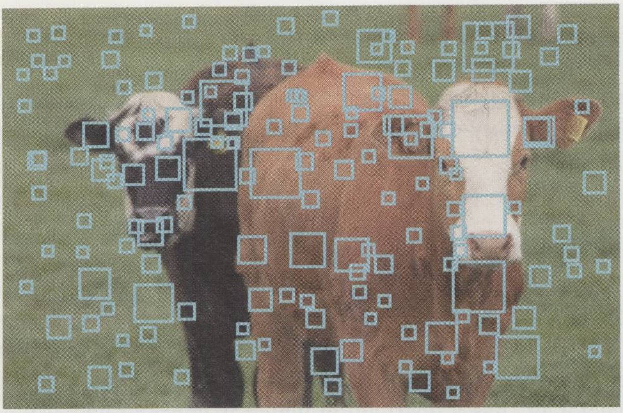

**Fig. 1.** Difference of Gaussian interest points with their local regions, in which the squares are centered at the interest points and the size of the squares indicates the scale of the interest points. The SIFT descriptors and colour features are obtained from these square patches Note that interest points fall both on the objects of interest (the cows) and also on the background.

it to a feature vector $\mathbf{h}_n$ and then assigns this vector to an object category. Here the assumption is that each image belongs to one and only one of some number $K$ of mutually exclusive classes. In [3] the classifier was based either on naive Bayes or on support vector machines.

Here we use a linear softmax model since this can be readily extended to determine feature saliency as discussed shortly. Thus the model computes a set of outputs given by

$$y_k(\mathbf{h}_n, \mathbf{w}) = \frac{\exp(\mathbf{w}_k^{\mathrm{T}} \mathbf{h}_n)}{\sum_l \exp(\mathbf{w}_l^{\mathrm{T}} \mathbf{h}_n)} \tag{2}$$

where $k \in \{1, \ldots, K\}$. Here the quantity $y_k(\mathbf{h}_n, \mathbf{w})$ which can be interpreted as the posterior probability that image vector $\mathbf{h}_n$ belongs to class $k$. The parameter vector $\mathbf{w} = \{\mathbf{w}_k\}$ is found by maximum likelihood using iterative re-weighted least squares [10]. We shall refer to this approach as VQ-S for vector quantized softmax. Results from this method will be presented in Section 6.

An obvious problem with this approach is that the patches which contribute to the feature vector come from both the foreground object(s) and also from the background. Changes to the background cause changes in the feature vector even if the foreground object is the same. Furthermore, some foreground patches might occur on objects from different classes, and are therefore provide relatively little discriminatory information compared to other patches which are more closely associated with particular object categories.

We can address this problem using the Bayesian technique of *automatic relevance determination* or *ARD* [8]. This involves the introduction of a prior distribution over the parameter vector w in which each input variable $h_c$ has a separate hyperparameter $\alpha_c$ corresponding to the inverse variance (or precision) of the prior distribution of the weights $\mathbf{w}_c$ associated with that input, so that

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{c=1}^{C} \mathcal{N}(\mathbf{w}_c|\mathbf{0}, \alpha_c^{-1}\mathbf{I}). \tag{3}$$

During learning the hyperparameters are updated by maximizing the marginal likelihood, i.e. the probability of the training labels $D$ given $\boldsymbol{\alpha}$ in which $\mathbf{w}$ has been integrated out, given by

$$p(D|\alpha) = \int p(D|\mathbf{w})p(\mathbf{w})\, d\mathbf{w}. \tag{4}$$

This is known as the *evidence procedure* and the values of the hyperparameters found at convergence express the relative importance of the input variables in determining the image class label. Specifically, the hyperparameters represent the inverse variances of the weights, and so a large value of $\alpha_c$ implies that the corresponding parameter vector $\mathbf{w}_c$ has a distribution which is concentrated around zero and so the associated input variable $h_c$ has little effect in determining the output values $y_k$. Such inputs have low relevance. By contrast a high value of $\alpha_c$ corresponds to an input $h_c$ whose value plays an important role in determining the class label. The inclusion of ARD leads to an improvement in classification performance, as discussed in Section 6. We shall refer to this model as VQ-ARD.

With this approach we can rank the patch clusters according to their relevance. The logarithm of the inverse of the hyperparameter $\alpha_c$ is sorted and plotted in Figure 2.

Equivalently this can be plotted as a histogram of $\alpha_c$ values, as shown in Figure 3. It is interesting to note that in this problem the hyperparameter values form two groups in which one group can loosely be considered as relevant and the other as not relevant, so far as the discrimination task is concerned.

Figure 4 shows the properties of the most relevant cluster and of the least relevant cluster, as well as that of an intermediate cluster, according to the ARD analysis based on $C = 100$ cluster centers. Note that the images have been hand segmented in order to

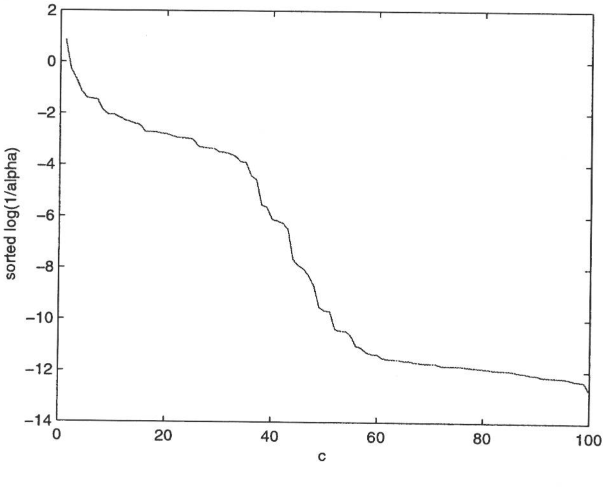

**Fig. 2.** The sorted values of the log variance (inverse of the hyperparameter $\alpha$)

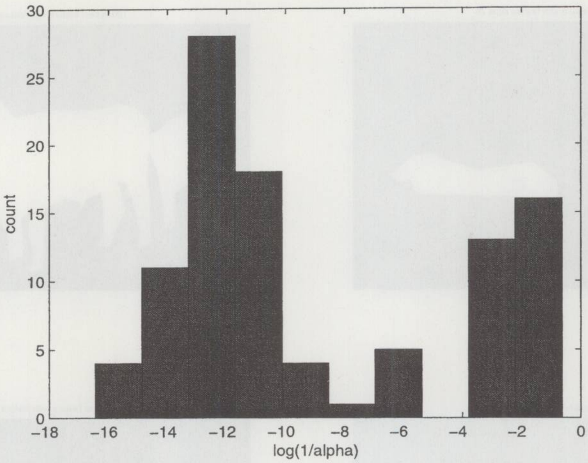

**Fig. 3.** The histogram of the log variances

identify the foreground region. This segmentation is used purely for test purposes and plays no role during training. The top row shows the features belonging to the worst cluster, i.e. ranked 100, on a sheep image and on a cow image. This feature exists in both classes and thus provides a little information to make a classification. The middle row shows the locations of patches assigned to the cluster which is ranked 27, in which we see that all of the patches belong to the background. Finally, the bottom row of the figure shows the features belonging to the most relevant cluster, ranked 1, on the same sheep and cow images. This feature is not observed on the sheep image but there are several patches assigned to this cluster on the cow image. Thus the detection of this feature is a good indicator of the presence of a cow.

It is also interesting to explore the behaviour of the two groups of clusters corresponding to the two modes in the distribution of hyper-parameter values shown in Figure 3. Figure 5 shows examples of cow and sheep images in each case showing the locations of the clusters associated with the two modes.

Although this approach is able to focuss attention on foreground regions, we have seen that not all foreground patches have high saliency, and so this approach cannot reliably identify regions occupied by the foreground objects. We therefore turn to the development of new models in which we explicitly consider the identity of individual patches and not simply their saliency for overall image classification. In particular the hard quantization of K-means is abandoned in favour of more probabilistic approaches. First we discuss a discriminative model and then we turn to a complementary generative model.

## 4 The Discriminative Model with Patch Labelling

Since our goal is to determine the class membership of individual patches, we associate with each patch $j$ in an image $n$ a binary label $\tau_{njk} \in \{0, 1\}$ denoting the class $k$ of

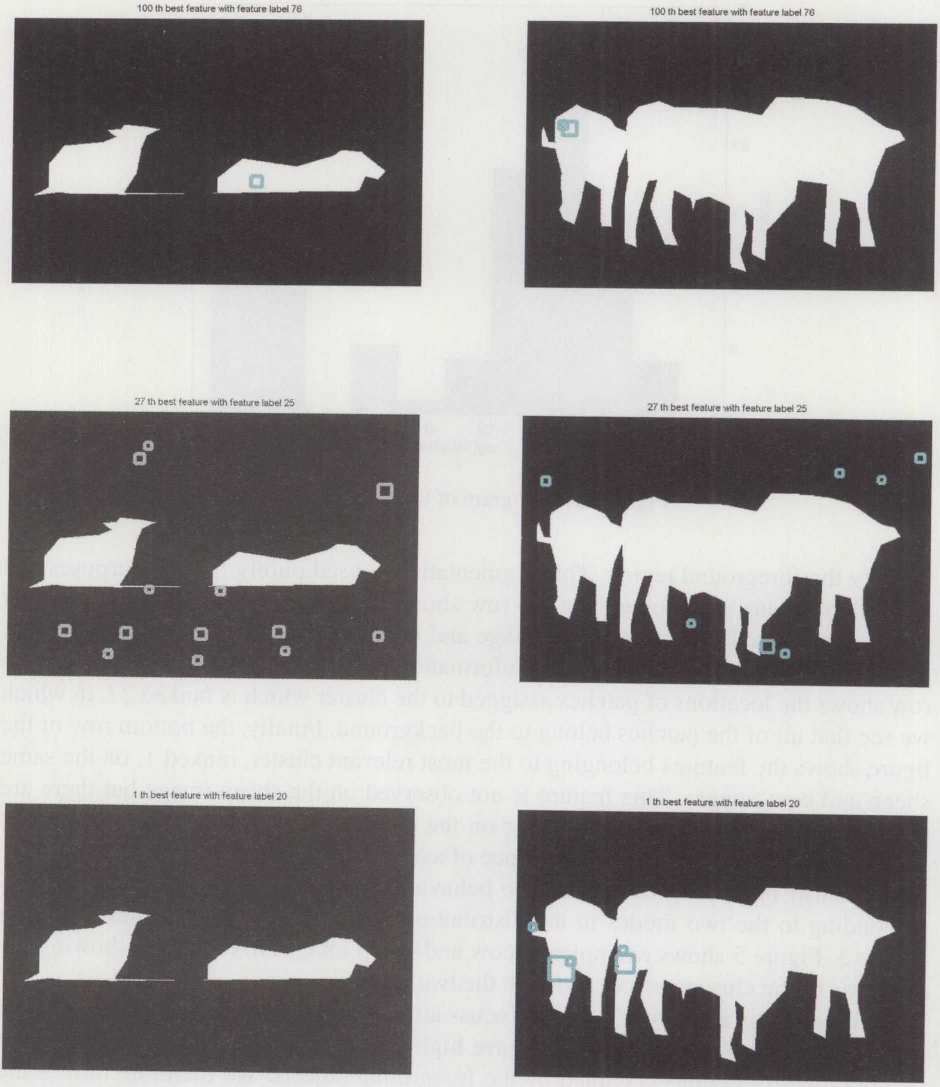
100 th best feature with feature label 76
100 th best feature with feature label 76
27 th best feature with feature label 25
27 th best feature with feature label 25
1 th best feature with feature label 20
1 th best feature with feature label 20

**Fig. 4.** The top row shows example cow and sheep images, with the foreground regions segmented, together with the locations of patches assigned to the least relevant (ranked 100) cluster center. Similarly the middle row analogous results for a cluster of intermediate relevance (ranked 27) and the bottom row shows the cluster assignments for the most relevant cluster (ranked 1). The centers of the squares are the locations of the patches from which the features are obtained and the size of the squares show the scale of the patches.