From Eigenspots to Fisherspots - latent spaces in the nonlinear detection of spot patterns in a highly varying background

Bjoern H. Menze, B. Michael Kelm, and Fred A. Hamprecht

Interdisciplinary Center for Scientific Computing (IWR),
University of Heidelberg,
INF 368,
69120 Heidelberg, Germany

from Multidimensional Image Processing , IWR, University of Heidelberg

From Eigenspots to Fisherspots – latent spaces in the nonlinear detection of spot patterns in a highly varying background

Bjoern H. Menze, B. Michael Kelm, and Fred A. Hamprecht

Interdisciplinary Center for Scientific Computing (IWR),
University of Heidelberg,
INF 368,
69120 Heidelberg, Germany

Summary. We present a scheme for the development of a spot detection procedure which is based on the learning of latent linear features from a training data set. Adapting ideas from face recognition to this low level feature extraction task, we suggest to learn a collection of filters from representative data that span a subspace which allows for a reliable distinction of a spot vs. the heterogeneous background; and to use a non-linear classifier for the actual decision. Comparing different subspace projections, in particular principal component analysis, partial least squares, and linear discriminant analysis, in conjunction with subsequent classification by random forests on a data set from archaeological remote sensing, we observe a superior performance of the subspace approaches, both compared with a standard template matching and a direct classification of local image patches.

1 Introduction – spot detection

In the hot and dry plains of ancient Mesopotamia and other parts of the Near East, but also in an arc stretching from the Balkans to India, small artificial mounds indicate the sites of early human settlements, some of them – as the biblic Jericho – being the remains of the first urban metropoles.

These so called "*tells*" are the result of millennia of settlement activity. Their base layers often reach as far back as 6000BC and a mud-based construction technique, prevalent to these regions, allowed some of them to raise up to significant heights during the millennia, forming characteristic landmarks. Though a large number of these mounds are well studied, the best current listings of them are neither comprehensive nor accurate. – However, in the digital elevation model of the Space Shuttle radar topography mission (SRTM), tells can be identified as small contrasting spots within the elevation pattern of the natural variation of the land surface [1].

As agricultural landuse and the growth of modern settlements impose an immanent threat to this cultural heritage and a study of the distribution of these former settlements is of high archaeological interest, we seek for a robust machine based processing of the SRTM data which allows for a fast, objective and precise guidance to tell sites in order to document them in wide regions of Turkey, Syria, Iraq and Iran.

Spot or point detection is a standard task in low level image processing. While elementary template matching is optimal for detecting point-like patterns in uncorrelated noise, other approaches exist in applications as diverse as preprocessing of microarray and gel electrophoresis image data [2, 3], the detection of cars in thermal bands of satellite imagery [4], or peak detection in 2D mass spectrometric data [5], to name a random selection. – Most of the spot detection approaches can be categorized into two classes: Parametric models are used to characterize the object, e.g. gaussian functions to model the spots, splines to fit and correct for the background. Alternatively, the detection is based on a phenomenological and nonparametric description of characteristic features, e.g. when searching for local extremes by morphological operations (watershed transformation), or evaluating the gradient images by structure tensors.

Unfortunately, a simple matched filter fails in the detection of tell-like mounds in the digital elevation model due to a high number of false positive hits. Also, the lack of positional a priori information, the variation of the spot pattern (diameter and height of the tell), and the highly variable "background", given by the natural topographic variation (ridges, walls, natural mounds), prohibit the application of spot detection algorithms as the ones mentioned above. –

Adapting ideas from face recognition, notably the concepts of *"Eigen"*- and *"Fisherfaces"* (see [6] and references therein), we learn adaptive templates (section 2) from our data (section 3.1), extending the idea of a (single) template matching to a multi-dimensional subspace approach for spot detection. Combined with a nonlinear classifier - random forests - we quantitatively compare (sections 3.2) and discuss (section 4) different methods intermediate between *Eigen*- and *Fisherspots* for our task.

2 Subspace filters – latent spaces

The optimal filter for the detection of a signal with known shape in additive white Gaussian noise is the *matched filter* (MF) [7]. Convolving an image with the MF can be regarded as correlating the image with a template of the signal to be detected. From a learning perspective, and extending the idea of a signal detection to a binary classification task between (tell) pattern vs. (non-tell) background, this approach corresponds to regarding the image as a collection of (local and independent) patches. All pixels in a patch are explanatory variables with an associated label, ie. pattern or background. In this feature space, the matched filter defines a one-dimensional linear subspace which is used to discriminate these two classes. From this point of view, the MF is very much related to linear regression methods, which motivates the approach taken in this paper and the naming *subspace filter*.

Real situations do not necessarily fulfill the ideal conditions under which the MF is proven to be optimal. Instead of seeking an optimal one-dimensional subspace and thus presuming linear separability in the feature space, we propose to perform a less restrictive dimensionality reduction, i.e. the projection onto a subspace of higher dimension followed by a nonlinear decision rule.

A common basic approach to the construction of a subspace which captures the most important variations in high dimensional data is *principal component analysis* (PCA). Its ranking criterion for the kth direction β_k is derived from the empirical covariance of the features :

$$\beta_{PCA_1,k} = \underset{\substack{||\beta||=1 \\ corr(\beta_j,\beta_k)=0, j<k}}{\arg\max} \quad var(X_1\beta) \tag{1}$$

with $corr(\beta_k, \beta_j)$ denoting the correlation between β_k and β_j; and where X_1 only holds the examples with the sought pattern. This projection compresses variation and information of the correlated spatial signal, but it neglects knowledge about the background signal X_0 and the binary character of the detection problem. In order to incorporate knowledge about X_0, PCA can be extended to derive the directions β_{PCA} from the variance of the full training data set X. This represents the prior belief that the variance of the training sample is due to interclass variations which are represented by the major eigendirections in the sample space.

The two-class information can be used explicitly as done in canonical correlation analysis (CCA). For univariate Y this is equivalent to ordinary least squares (OLS) regression [8] which, for the two-class problem, yields the same directions as linear discriminant analysis (LDA) [9, p.88]. All these problems determine the optimal direction β based on the correlation between the class label Y and the projected feature scores $X\beta$. They choose directions with high discriminative power:

$$\beta_{LDA,k} = \underset{\substack{||\beta||=1 \\ corr(\beta_j,\beta_k)=0, j<k}}{\arg\max} \; corr^2(X\beta, Y) \tag{2}$$

again with orthogonal directions β_k for linearly nonseparable problems. – OLS and LDA are known to have bad generalization performance in the presence of collinear features, i.e. they are vulnerable to overfitting (e.g. see OLS projections in fig. 6).

Introducing a bias, forcing subspace projections to more "realistic" directions with higher data support, can help to overcome this problem. Regularization is obtained by combining the two strategies mentioned above and optimizing for covariance or equivalently for the product of variance and squared correlation [10]:

$$\begin{aligned}
\beta_{PLS,k} &= \underset{\substack{||\beta||=1 \\ corr(\beta_j,\beta_k)=0, j<k}}{\arg\max} \; cov^2(X\beta, Y) && (3) \\
&= \underset{\substack{||\beta||=1 \\ corr(\beta_j,\beta_k)=0, j<k}}{\arg\max} \; corr^2(X\beta, Y)\, var(X\beta) && (4)
\end{aligned}$$

This forces the directions of the subspaces to have a natural “backing” in the data variation: the solution is pulled away from the OLS solution of maximal correlation towards directions of maximal variance in sample space as obtained by PCA.

Two related methods allow to vary the influence of the variance continuously. Ridge regression/penalized discriminant analysis (RR/PDA) extends the concept of OLS/LDA [10]:

$$\beta_{RR,k}(\gamma) = \underset{\substack{||\beta||=1 \\ corr(\beta_j^T \beta_k)=0, j<k}}{\arg\max} \quad corr^2(X\beta, Y)\,\frac{var(X\beta)}{var(X\beta)+\gamma} \tag{5}$$

A generalization of PLS is continuum regression (CR) [11]:

$$\beta_{CR,k}(\gamma) = \underset{\substack{||\beta||=1 \\ corr(\beta_j, \beta_k)=0, j<k}}{\arg\max} \quad corr^2(X\beta, Y)\,var(X\beta)^{\gamma} \tag{6}$$

Both approaches come at the cost of a hyperparameter γ to be tuned in addition to the optimal subspace dimension λ. Because of this and since PLS provides means to regularize LDA they will not be studied in the following.

3 Methods

3.1 Data

Tell sites. Average tells reach a height of 10-50m and have a diameter of 50-500m. In the SRTM elevation data set their patterns appear as small bright spots of one to five pixels diameter and with approximate radial symmetry (cf. fig. 1). In the SRTM of a North Syrian plain, the Khabur basin [12], positions of 184 known tell sites could be identified. In addition, 50 000 locations were randomly sampled (with uniform distribution) from the same geographic region as representatives of the background class X_0. An independent test data set, comprising positions of another 133 sites, was available from an archaeological survey in the same area [13].

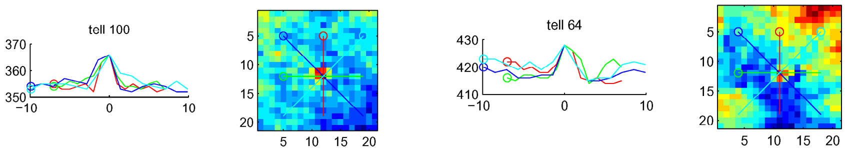

Fig. 1. Point patterns in the digital terrain model. Left: Profiles. Right: Top view, profile sections indicated.

Features. Elevation data from circular regions of 1km diameter, centered around the training sites, was used as input for the classifier design (compare geometry of resulting filters: fig. 6). To remove the absolute elevation, the feature vector contained height differences relative to the center of the image patch. The spatial extensions of the patch and therefore the optimal scale of the detection problem were assessed from the random forest Gini importance ($P = 80$, fig. 2). Rotational symmetry was assumed for the tell pattern. Accordingly, tell patterns rotated by 90, 180 and 270 degrees were also included in the training set, increasing the number of data points within X_1 to $N_1 = 736$.

3.2 Benchmark

The performance of a number of filters were compared quantitatively: PCA on the event class (PCA_1), PCA, MF, LDA and PLS on both classes (see table in fig. 3). The subspace scores of these filters were used for learning of the following multivariate decision rule.

Random forest [14] was chosen as decision rule on the various filter responses and was also applied to the original data without intermediate dimension reduction. Random forest models the posterior probability of a class by an ensemble of trees on bootstrapped data sets. In contrast to traditional bagging, only a limited number of features is randomly chosen in the search for the optimal split at each node. Its advantage is the ease and speed of training, while its performance is comparable to other state of the art classifiers, such as support vector machines.

In the error estimation, a tenfold cross-validation over a predefined spatial grid of 60 non-overlapping boxes ($15^2 km^2$ each, covering the Khabur basin) was chosen due to the spatial correlation of the data. Before applying them to the holdout data, filter and classifier were optimized via a fivefold inner cross-validation loop, also over the spatial grid. Within this step, the subspace dimensionality was increased from $\lambda = 1, ..., 10$, while the classifier settings were kept unchanged (300 trees, one randomly chosen variable at the nodes).

For the error quantification, the area under curve of the receiver operator characteristic (ROC AUC) was used to provide an integrated measure of sensitivity (true positives / all positives) and specificity (true negatives / all negatives). In the final evaluation also precision (true positives / (true positive + false positives)) and recall (= sensitivity) were considered, since these measures focus on the event class.

4 Results and Discussion

Both PCA and PLS result in filter sets whose first component are similar to a *matched filter* (fig. 6), hence their higher components indeed can be seen as higher dimensional extension to a MF. The performance of the one dimensional MF (see table 3) is exceeded by any multidimensional filter approach,

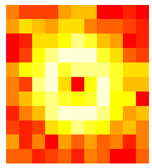

Fig. 2. Relevant features in the classification of spots and background. The size of the image patch and filter mask are determined by the random forest Gini importance (ranked, red/yellow – low/high importance). The central pixel is constant zero for all samples, see text.

Threshold	PCA	PCA$_1$	PLS	LDA	MF	all
FN 0.9	7.2	8.2	**7.1**	13.0	36.5	20.7
FN 0.95	5.0	4.9	**4.6**	11.0	33.7	16.7
FN 0.99	2.6	**1.5**	2.6	6.9	25.3	13.6
FP 0.9	6.7	7.4	**6.1**	8.8	13.5	7.6
FP 0.95	13.0	14.0	**12.0**	**12.0**	17.0	12.8
FP 0.99	57.0	59.0	48.0	**22.0**	23.4	31.5

Fig. 3. Table: Classification accuracy for different thresholds. False negatives (FN) in % of the target class, false positives (FP) in $\%_0$ of the background class (compare to fig.4).

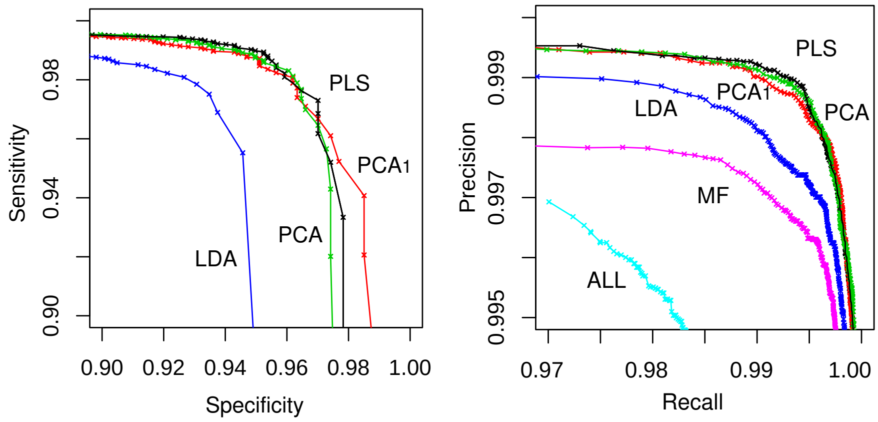

Fig. 4. Classification performance: receiver-operator-characteristic (left), precision-recall-curve (right). "ALL" denotes the direct application of the classifier without subspace filter.

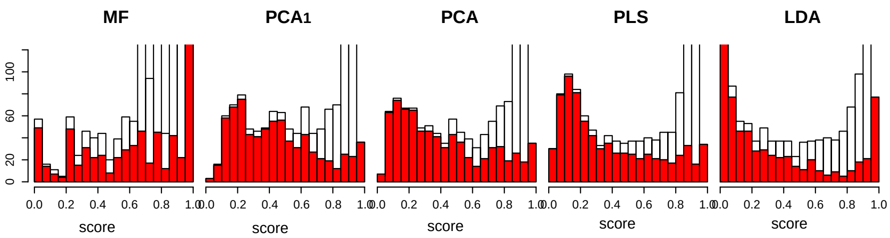

Fig. 5. Distributions of the event signals (red) and the background (white) from the test data (random forest probability). Histograms are truncated, the total number of counts is 50736.

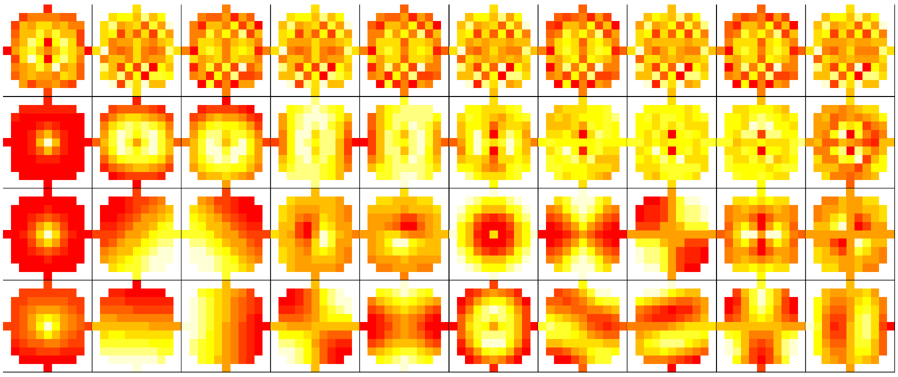

Fig. 6. First ten subspace filters for LDA, PLS, PCA_1, PCA (from top to bottom).

while the direct application of the non-linear classifier to unfiltered data leads to a classification performance surpassed by *any* subspace approach. During resampling, the optimal dimensionality of these filters λ was between 5 and 7.

The application of *linear discriminant analysis* results in a distinct separation of the data and a nearly binary distribution of the scores (fig. 5). Falsely classified signals also appear at the tails of the distribution, thus leading to the weak performance of LDA under the ROC and the precision-recall curve. The oscillating checker-board patterns in the filter set (fig. 6) indicate an overfitting on the highly collinear image data, explaining the comparably bad generalization behavior (table 3).

Principal component analysis performs very well in both variants (PCA, PCA_1). The distribution of the scores (fig. 5) shows a higher variance than both PLS and LDA. The orthogonal loadings of PCA_1 are adapted to variants of the central point pattern, while loadings of PCA explain the overall variation (fig. 6) in the data set. Classification in the PCA subspace controls false positives better than in the PCA_1 subspace (table 3), while the latter allows the highest specificity/recall (fig. 4) of all methods at the cost of a somewhat lower overall precision.

The shape of the *partial least squares* feature distribution is in-between the distribution of LDA (max. correlation) and PCA (max. variation), reflecting the intermediate character of PLS. On the present data, PLS is optimal under the precision/recall curve (fig. 4) and in the control of false positive events, although the differences between PLS and PCA remain faint. –

In our data set, PCA filters obtained from both classes perform nearly as well as PCA filters learned only from the spot class (PCA_1). Based on our experience with similar problems, we argue that this a special feature of the present data set, while in general a good performance of the (two-class) PCA crucially depends on the appropriate choice of the background samples. Accordingly, we recommend to apply PCA_1 if a highly precise representation of the (spot-) pattern is sought and to consider PLS if the use of both classes and an explicit incorporation of background prototypes is desired in the definition of the subspace filters.

While the complementary concepts of Eigen- and Fisherfaces (PCA, LDA) are the most frequently applied in face recognition, we can observe an advantage of the regularized subspace filters (PCA, PLS) on our local image patches, setting the presented low level feature extraction in proximity to chemometrical data analysis rather than classical image processing. We note that the definition of the relevant scale in our detection problem – the extensions of the local image patches – by the multivariate random forest importance is novel.

Applying the PLS filter on the digital elevation model of the geographical region with the available archaeological ground truth [12], it is possible to detect all (regular) settlement mounds higher than 5-6m (85/133) with 327 false positives in a tile of 600*1200 pixels. This allows us to use the presented spot detector in a screening of wide regions of the Near East and for a joint, machine based evaluation with other remote sensing modalities.

5 Conclusions

Extending the idea of a matched filtering (to be followed by a threshold operation) to the training of higher dimensional latent space filters combined with a subsequent nonlinear classifier proves to be a viable concept in the presented spot detection. If a (binary) training data set is available, this approach can be the appropriate choice for a detection of spot patterns in a highly varying background, supporting or replacing traditional parametric spot detectors.

Acknowledgement

BM gratefully acknowledges financial support by the MFG Steinbuch award.

References

[1] Sherratt, A; Spotting tells from space; Antiquity 77, 2004

[2] Rahnenfuehrer, J and Bozinov, D; Hybrid clustering for microarray image analysis combining intensity and shape features, BMC Bioinformatics 5:47, 2004

[3] Boetticher, GD et al.; A recursive application of a support vector machine for protein spot detection in 2-dimensional gel electrophoresis, Journal of Computer Science 1:355-362, 2005

[4] Hinz, S; Fast and subpixel precise blob detection and attribution, Proceedings of ICIP, 2005

[5] url: http://arep.med.harvard.edu/MapQuant/; last access: 01. April. 2006

[6] Belhumeur, PN et al.; Eigenfaces vs. Fisherfaces: Recognition using class specific linear projections, Proceedings of ECCV, 1996

[7] Moon, TK and Stirling, WC; Mathematical Methods and Algorithms for Signal Processing; Prentice Hall; New York, 2000

[8] Borga, M, A unified approach to PCA, PLS, MLR and CCA; techn. report; Linkoping, Sweden, 1997

[9] Hastie, T et al., The Elements of Statistical Learning, Springer, New York, 2001

[10] Frank, IE and Friedman, JH; A statistical view of some chemometric regression tools; Technometrics 35:109-148, 1993

[11] Bjorkstrom, A; A generalized view on continuum regression; Scandinavian Journal of Statistics 26:17-30, 1999

[12] Menze, BH and Ur, JA and Sheratt, AG; Detection of ancient settlement mounds: Archaeological survey based on the SRTM terrain model; Photogrammetric Engineering & Remote Sensing 72:321327, 2006

[13] Ur, JA. Settlement and landscape in Northern Mesopotamia: The tell Hamoukar survey 2000-2001, Akkadica 123:5788, 2002

[14] Breiman, L.; Random forests; Machine Learning 45:532, 2001