Learning Linear Discriminant Projections for Dimensionality Reduction of Image Descriptors

Hongping Cai[1,3] Krystian Mikolajczyk[1] Jiri Matas[2]
[1]University of Surrey, UK [2]CTU Prague, Czech Republic
[3]National University of Defense Technology, China
{h.cai,k.mikolajczyk}@surrey.ac.uk matas@cmp.felk.cvut.cz

Abstract

This paper proposes a general method for improving image descriptors using discriminant projections. Two methods based on Linear Discriminant Analysis have been recently introduced in [3, 11] to improve matching performance of local descriptors and to reduce their dimensionality. These methods require large training set with ground truth of accurate point-to-point correspondences which limits their applicability. We demonstrate the theoretical equivalence of these methods and provide a means to derive projection vectors on data without available ground truth. It makes it possible to apply this technique and improve performance of any combination of interest point detectors-descriptors. We conduct an extensive evaluation of the discriminative projection methods in various application scenarios. The results validate the proposed method in viewpoint invariant matching and category recognition.

1 Introduction

Local image descriptors have become a strategy of choice for addressing a wide variety of problems in computer vision, from wide baseline matching and the recognition of specific objects to the recognition of object classes. They have been applied to image retrieval, panorama building, texture recognition, scene classification, robot navigation, visual data mining etc. The most widely used descriptor is SIFT [8] due to its simplicity, robustness to common image perturbations and excellent performance in finding visual correspondences between images. It represents state-of-the-art and many of its variants have been reported in the literature [4, 10], however these are mainly manually designed modifications to strike different performance trade-offs. A different strategy was introduced in [14]. The idea is to break the feature extraction process into simple and parameterized modules and then learn the optimal parameters of descriptors to maximize their matching performance. Extensive experimental evaluation reported in this paper indicates the trade-offs and the best set of descriptor parameters for matching images from different views. It is however inconvenient to perform such optimization in every possible application scenario. It requires large training set with a ground truth which is not always available or straightforward to establish. Furthermore the resulting descriptors are high dimensional which is one of its main issues in the context of large scale experiments. There are however effective solutions for similarity search in low dimensional spaces [8, 11]. Standard PCA has been widely used in the community to reduce the number of dimensions [4, 10]

but the performance often dropped as well. Recently, a method for dimensionality reduction based on Fisher Analysis has been applied to local descriptors [3, 11]. It leads to better results than PCA as it gives a set of projections that maximize inter to intra cluster distances. Improved matching performance has been demonstrated in [3] as well as more efficient search in various tree like data structures in [11]. Consistent improvements in different test scenarios indicate that the method generalizes well which is crucial given the wide range of applications of local descriptors. There is however an important drawback which makes this approach less attractive. It requires training on large dataset with ground truth. A large set of correctly matched features is used to estimate intra-class covariance. While in wide baseline matching it is possible to establish unique correspondences by applying geometric constraints, it is much more difficult to find such correspondences in object class recognition. The rigid transformations cannot be used here and very similar features occur in different positions on objects of the same category. We also observed that there are no optimal general discriminant projections for various data and various application. The projections have to be generated through subset of data from a given application to improve descriptor performance. It is crucial to perform it without using ground truth.

The main contribution of this paper is an evaluation of two discriminant projection techniques for dimensionality reduction as well as a strategy for learning projection vectors, which does not require data with a ground truth. We demonstrate theoretical equivalence of the two approaches recently proposed in [3] and in [11] and investigate the practical differences. The proposed dimensionality reduction outperforms PCA, it is applicable to various application scenarios and straightforward to implement. We investigate different regularization methods for providing stable projection vectors. Finally, we perform extensive experimental evaluation to demonstrate that the method brings improvement for different applications.

2 Linear discriminant projections

In this section, we give an overview of linear discriminant projections defined in [3] and [11]. The aim is to project a feature descriptor so that the projected corresponding features get closer while the projected different features get farther in projected feature space. We define X to be the feature space and l_{ij} to label feature correspondences. $l_{ij} = 1$ indicates that x_i and x_j ($x_i, x_j \in X$) are matched features, while $l_{ij} = 0$ indicates unmatched features.

In [3], a projective direction W is designed to maximize the ratio of variance between unmatched features and matched features. The solution is the generalized eigenvectors:

$$W = eigenvectors^T(B^{-1}A) \tag{1}$$

where A and B are covariance matrices of matched feature differences and unmatched features respectively: $A \stackrel{\text{def}}{=} \sum_{l_{ij}=0}(x_i - x_j)(x_i - x_j)^T$, $B \stackrel{\text{def}}{=} \sum_{l_{ij}=1}(x_i - x_j)(x_i - x_j)^T$.

In [11], a linear projective transformation applied to SIFT and termed M-SIFT includes two parts. The first is the inverse of the square root of intra-class covariance matrix, which is used for whitening of the original feature space. The second part is PCA of the unmatched features in $\tilde{Y}$ space ($\tilde{Y} = \{B^{-\frac{1}{2}}x | x \in X\}$):

$$P = eigenvectors^T(B^{-\frac{1}{2}}AB^{-\frac{1}{2}}) \cdot B^{-\frac{1}{2}} \tag{2}$$

In the remainder of this paper, W and P are used to refer to the linear discriminant projections presented in [3] and [11], respectively.

2.1 Equivalence

Although different methods are used in [3] and [11] to obtain the projective vectors, they have the same purpose which is to find an optimal projective orientation to separate the matched features from unmatched features. They also use the same information, which is the covariance matrices of matched feature differences and unmatched features as shown in equations (1) and (2). We demonstrate their equivalence as follows:

Let R be a matrix of all eigenvectors of $B^{-\frac{1}{2}}AB^{-\frac{1}{2}}$ sorted by eigenvalue magnitude and let Ω be a diagonal matrix of the corresponding eigenvalues:

$$B^{-\frac{1}{2}}AB^{-\frac{1}{2}} \cdot R = R \cdot \Omega \tag{3}$$

Equation (3) can be multiplied by $B^{-\frac{1}{2}}$, we obtain $B^{-\frac{1}{2}} \cdot B^{-\frac{1}{2}}AB^{-\frac{1}{2}}R = B^{-\frac{1}{2}} \cdot R\Omega$, hence:

$$B^{-1}A \cdot B^{-\frac{1}{2}}R = B^{-\frac{1}{2}}R \cdot \Omega \tag{4}$$

where $B^{-\frac{1}{2}}R$ is the eigenvectors of $B^{-1}A$, which means that $W = eigenvectors^T(B^{-1}A) = (B^{-\frac{1}{2}}R)^T = R^TB^{-\frac{1}{2}} = P$.

The equivalence of projections P and W is also verified by the theorem of simultaneous diagonalization of the covariance matrices. Let Y^P be the projected feature spaces: $Y^P = \{Px|x \in X\}$.

$$B|_{Y^P} = cov(P \cdot (x_i - x_j))_{l_{ij}=1} = P \cdot B \cdot P^T = RB^{-\frac{1}{2}}BB^{-\frac{1}{2}}R^T \tag{5}$$

Since R is an orthogonal matrix, $B|_{Y^P} = I$.

$$A|_{Y^P} = cov(P \cdot x_i)_{i \in \{l_{ij}=0\}} = P \cdot A \cdot P^T = RB^{-\frac{1}{2}}AB^{-\frac{1}{2}}R^T \tag{6}$$

R is the matrix of all eigenvectors of $B^{-\frac{1}{2}}AB^{-\frac{1}{2}}$, thus $A|_{Y^P} = \Omega$.

From the above equations, P is a linear transformation that diagonalizes two symmetric matrices B and A simultaneously. According to the theorem in [1](p.31), Ω and P are the eigenvalue and eigenvector matrices of $B^{-1}A$, which demonstrates that P equals W. Experimental results validate this equivalence. Figure 1 shows P and W for NG (normalized grayvalue patches) and for SIFT features. However, one can notice a difference between P and W in figure 1, for example in the third column of NG patches or displayed vectors of SIFT. The magnitudes of the projective vectors differ. Hence P and W have the same projective orientations only. If we decompose B: $B = \phi\Lambda\phi^T$, we can express projection P as:

$$P = eigenvectors^T(B^{-\frac{1}{2}}AB^{-\frac{1}{2}}) \cdot \Lambda^{-\frac{1}{2}}\phi^T \tag{7}$$

According to equations (7) and (1), P not only rotates the feature space but also scales the dimensions, while W only rotates the coordinates. If P_i and W_i are defined the i-th projective vectors of P and W, then $\|P_i\| \neq 1$ since $\Lambda \neq I$, while $\|W_i\| = 1$ as these are eigenvectors. The orientation defines the projection, therefore in theory descriptors projected with P_i and W_i should give the same performance. However, in practice, we observe different performances in particular in high dimensional spaces. This is due to the insufficient data for estimating covariance matrices, hence the smaller eigenvalues of B are unreliable. A little variation of small eigenvalues λ_{small} of B brings on large changes to $1/\sqrt{\lambda_{small}}$, which results in incorrect scaling of the feature space with projections P. As we demonstrate in our experiments, high dimensions of P-projected features are unreliable. On the contrary, W does not scale the feature space, so the dimensions corresponding to small eigenvalues are less affected. This property is verified in the patch matching experiments in section 3.1 .

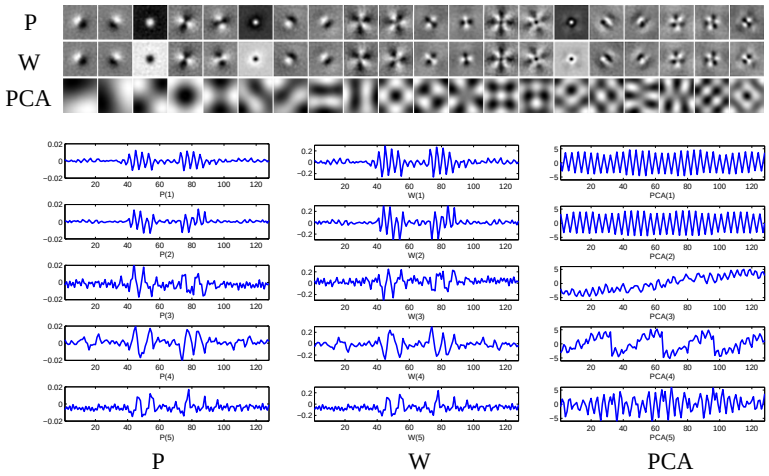


Figure 1. Projective vectors of P, W and PCA. Top: The first 20 projections of normalized grayvalue feature. Bottom: The first 5 projections of SIFT feature. 69,000 matched pairs from [13] are used for estimating the projections.

2.2 Covariance estimation

A crucial problem in computing P or W is how to estimate the intra-class covariance matrix B. In this section we introduce different methods for estimation and regularization of this matrix and discuss the impact they may have on the projection vectors. In addition to estimation from ground-truth data, we explore two ways to estimate the intra-class covariance matrix without the ground-truth.

Ground-truth data. In [3] and [11], the linear discriminant projections worked on the premise that a large amount of ground-truth data are used for intra-class covariance estimation. Matched feature pairs are produced by bundle adjustment in images of 3D scenes [13] and homography of image pairs [10]. However, for estimating the intra-class covariance matrix, problem of insufficient data may occur in high dimensional feature spaces. In order to tackle this problem, two regularization methods are investigated.

Power Regularization is one of solutions to insufficient training data. The idea is to replace a number of the smallest eigenvalues of B, with their maximum eigenvalue. The fraction of replaced eigenvalues is controlled by parameter α. This approach was used in [3] for high dimensional features. Figure 2 shows the effect of power regularization on NG feature. The projections are very noisy if no regularization is applied ($\alpha = 0$, top row in Figure 2). When α approaches 100% then B becomes an identity matrix multiplied with a constant. In this case, W and P degenerate to PCA: $W = P = eigenvectors^T(A)$ (see the bottom row in Figure 2 and PCA in Figure 1).

An alternative covariance regularization, often used in classification and recognition in high dimensional spaces, is to combine the inter-class covariance matrix with the intra-class matrix [12]: $\hat{B} = \alpha B + (1-\alpha)A + \beta I$, where α controls the mixing proportions. This method however was marginally successful in our experiments. It seems the unmatched feature covariance introduces incorrect bias and the discriminative character of the projections is reduced.

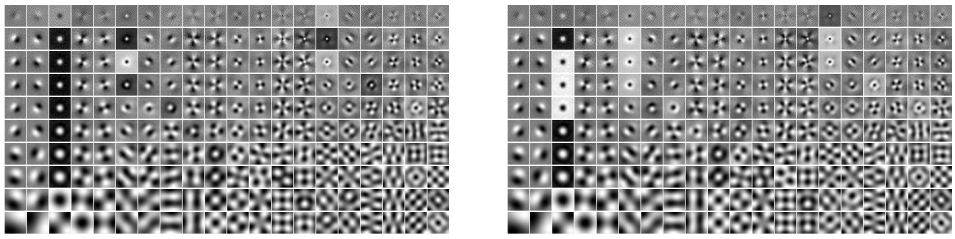


Figure 2. The first 20 projections of power regularized P and W for the normalized feature. Left: P, Right: W(From top to bottom, α=0, 10%, 20%,...,90%). The same setting as in Figure 1.

Simulated data. In order to approach the problem of insufficient or no ground-truth data, we investigate possibilities of generating it artificially. The idea is to simulate image transformations from detected image patches. All the transformed patches produced from the original one are considered matched. For simplicity, we only consider the affine transform and image blurring. We first employ single transformations to find the optimal parameters. Then we combine the transformations to simulate the real changes. We also show that this approach is successful for category recognition even though it is impossible to model real image deformations with parametric transformations (cf. section 3.3).

Clustered data. An alternative way to generate matched patch pairs for covariance estimation is to cluster all patches without ground truth. The features in the same cluster are considered as matched although in fact they might not be. We assume that nearest neighbors based on the descriptor similarity are correct matches and can provide sufficient information to estimate the covariance. It is a strong assumption but it may lead to better results in particular when used in combination with the simulated data. We employ the agglomerative clustering [6] to obtain the feature clusters. It is a bottom up clustering approach which iteratively finds the nearest features in the set and merges them into clusters. In our implementation features whose distances are smaller than a threshold are linked to a cluster. Smaller threshold results in less false matches for intra-class covariance, while larger threshold can produce more matched feature pairs. We select a threshold according to distance distributions of real matched and unmatched features estimated on ground-truth data (cf. figure 3).

3 Experiments

We present three main experiments on normalized image patches and SIFT descriptors. The first experiment is measuring patch matching performance to compare the projections obtained from the three training methods discussed in the previous section. Second experiment is testing the performance in the context of wide baseline matching. Finally, we incorporate the proposed method into a scene recognition system. In all experiments the performance of low-dimensional P- and W-projected features is compared with that of PCA projected ones and the original SIFT features.

3.1 Patch matching

In patch matching tests, both the training and testing data are image patches from Photo Tourism reconstructions [13], with associated ground truth. We randomly choose 69,000 matched patch pairs for training and 15,000 for testing. Note that there is no intersection between them. The nearest neighbor matching strategy is then applied to test the methods. To obtain the grayvalue features we sample the original patches to 32x32 pixels with bilinear interpolation. Then all the patches are normalized to 0-mean and 1-variance to reduce the effect of illumination changes. The SIFT descriptors are computed on unnormalized patches directly.

Tests on ground-truth data. We first estimate the distance distributions for matched and unmatched features as well as for their projections with P, W and PCA, which are shown in Figure 3 for SIFT features. The ratio of the intersection areas between matched and unmatched features for the original and the projected space indicates the discriminability of the projection. As is shown on the top of the figures, the intersection areas of projections P (0.021) and W (0.019) are smaller than those of PCA (0.024) and the original feature (0.022). Similar observations were made in experiments with grayvalue patches.

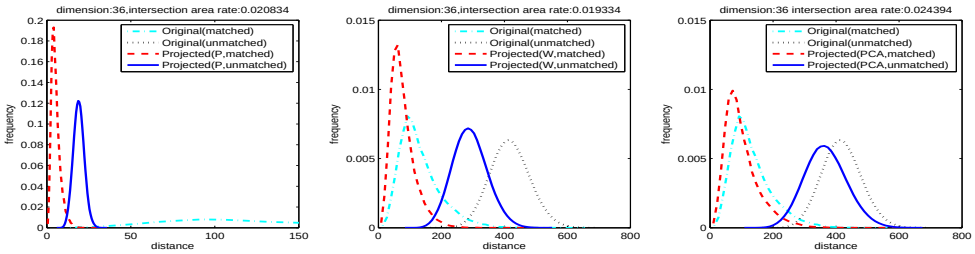
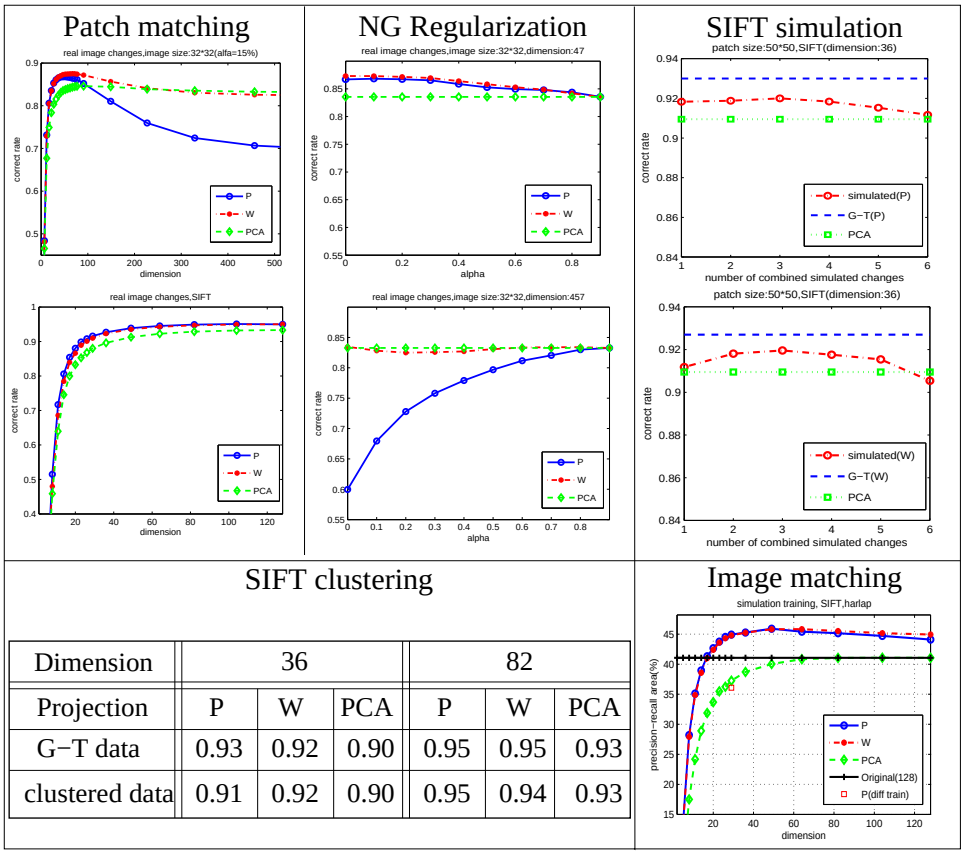


Figure 3. SIFT distance distribution: (left) P, (middle) W, (right) PCA.

The nearest neighbor matching performance of projected NG feature and SIFT w.r.t. dimensionality is shown in Figure 4(Patch matching). Compared with PCA, projections P and W improve the correct rate of NG by 4.5% for 32 dimensions and by 3.6% for SIFT feature with 29 dimensions. For SIFT, the performance of the projections increases with more dimensions, while for NG the correct rates of P and W decrease after a highest performance at 30 dimensions. This dropping performances is due to insufficient training data which is a problem for high number of dimensions. We observe that the overfitting problem is more significant for projection P. As we discussed in section 2.1, poor performance of P in high dimensions results from incorrect scaling of the feature space by small eigenvalues of P.

Power regularization of the covariance matrices (cf. section 2.2) is often used to overcome the problem of insufficient data and the matching performance for this method is presented in Figure 4(NG Regularization). Power regularization has little effect on W for both low and high number of dimensions. As to P, it little affects the performance of small number of dimensions (47), while there is significant improvement for large number of dimensions (457). When strong regularization is used, the matching performances of P and W are equal to that of PCA (cf. section 2.2). In practice, small number of top dimensions are used only, hence the power regularization is less important.

Tests on simulated data. In this test we randomly pick 23,000 unmatched patches and 5,000 matched patch pairs for training and testing. The training and testing sets have no intersection. We have tested different ways of generating patches and the best results



Dimension	36			82		
Projection	P	W	PCA	P	W	PCA
G−T data	0.93	0.92	0.90	0.95	0.95	0.93
clustered data	0.91	0.92	0.90	0.95	0.94	0.93

Figure 4. Discriminant projection and dimensionality reduction results for normalized grayvalue patches and SIFT descriptors.

were obtained when each patch was transformed with a random transformation. We use affine transformation decomposed to translation, rotation, scaling, skewing and squeezing. For each patch, four randomly selected transformations were applied with random parameters from an arbitrarily set range of values, hence there are $23,000 \times 4$ matched feature pairs for intra-class covariance matrix. Figure 4(SIFT simulation) displays the results for combinations of the individual transformations. The simulated P and W perform slightly better than PCA but are still lower than those of ground truth data (G-T). As shown in the figure, the differences between various combinations are small, in which the three-combination including scaling (1.05), skewing(0.06) and squeezing(1.1) produce the highest performance of 0.92 for both P and W. In the following we explore a different strategy to model nonlinear changes in real images.

Tests on clustered data. An alternative method for finding matched features is to use small similarity distance where the fraction of incorrect matches is very small. We use agglomerative clustering [6] to group the features into matched clusters. We select a threshold value for maximum cluster size from the intersection of two distributions of the matched and unmatched feature distances (see figure 3). Figure 4(SIFT clustering) illustrates the results of clustering training. Though the clustering data cannot reach the

performance of the ground-truth data, there is 1%-2% improvement with P and W compared to PCA. Another strategy which is to combine simulating and clustering ideas. After clustering the original features, we simulate image changes for each cluster member. The 'cluster+sim' achieves a small improvement of 2% over the ground truth data.

The results presented in this section confirm the observations from [3] although different performance measure was used there and the scores are not directly comparable. In our experiments however discriminative projections do not bring large improvements on the data from [13]. A possible explanation is that the original descriptor already matches correctly over 90% of the data and the remaining feature pairs are outliers which are difficult to model with linear transformations.

3.2 Image matching

In this section, we apply the discriminant projections to image matching. We test on image sequences from the publicly available set [10] with homography ground truth. The local regions are first extracted from a pair of images by Harris-Laplace detector which achieved high performance in [10]. We then estimate the intra and inter-class covariance matrices with the matched features produced by simulation and clustering of the detected regions.

We adopt the evaluation criteria from [10] which is the area below a precision-recall curve. The nearest neighbor matching is used to find the matched features in each image pair, and verified with the ground-truth homography.

Tests on simulated data. From the previous experiment, the best three simulated transformations are scaling(1.05), skewing(0.06), squeezing(1.1) and in addition we add rotation(5) as 'sim1'. Each transformation produces 4 patch pairs, which results in 16 matched feature pairs for each region. Figure 4(Image matching) shows the average matching performance with respect to the number of dimensions used. The original SIFT, indicated by the horizontal black line, serves as a reference. When at least 50 dimension is used, the performance of PCA projected features reaches that of the original SIFT. In contrast, P and W yield higher precision-recall area than PCA and also start to outperform the original 128-dim SIFT with 20 dimensions only. The performance of P and W reaches maximum after 29 dimensions. Harris-Laplace detector with 29-dim P-projected SIFT yields 3.9% higher precision-recall area than the original SIFT and 7.7% higher than 29-dim PCA-projected SIFT. We also experiment with the MSER detector [9], and we observe the improvement from P and W(dimension: 29) by 2.6% and by 6.4% compared with 128-dim SIFT feature and 29-dim PCA-projected feature.

Tests on clustered data. The clustering method is adopted in the training stage in a similar way to the experiment from section 3.1. All the images from each test sequence from [10] are used to extract features, form clusters and estimate the projections. 29-dim projections P and W improve the precision-recall area by 4.6% and 9.1% compared to the original and to the PCA-projected SIFT, respectively.

Generalization. We carry out another experiment with projection vectors trained for Harris-Laplace and then used with MSER for matching. The P and W perform similar to PCA and worse than the original features since the projections are not adapted to the data. Furthermore, the matching test is performed with projections obtained from the ground-truth data in section 3.1. Unfortunately the P-projected feature performs even worse than the PCA-projected feature, which is shown as a square in Figure 4(Image matching). These experiments show that the discriminant projections adapt to the data however we

found no general projections that would improve matching performance on any type of data. Consequently, our strategy without ground-truth is crucial for the applicability of discriminant projections.

3.3 Scene Recognition

To demonstrate the properties of the proposed method in different application scenario we perform scene recognition test using a system similar to the one proposed in [2]. Given a set of labeled training images the systems extract interest points, computes descriptors, from which it constructs a pyramid codebook with a kmeans clustering. It then collects occurrences of codewords on the training data and trains one-versus-all SVM classifiers. Given a query image the features are extracted, compared to the codebook and classified with the trained SVM. In our implementation the features are extracted and we estimate the projection vectors with simulating and clustering methods. We compare the results with the original SIFT and the PCA reduced descriptor.

We experiment with 15-class database from [5, 7]. The experiments are repeated 10 times with different randomly selected training images and testing images. In each class, 100 images are used for training and the rest for testing. To reduce computational complexity, only 5% of all regions are used to estimate projections with the simulating strategy. Figure 5 displays the comparison of the average recognition rates of original SIFT feature, PCA- and *P*-projected features. We also experimented with different simulating combinations which gave similar recognition rates, all higher than the original SIFT feature and PCA, while four clusterings produce lower performance. Increasing the clustering threshold slightly improves the score, which suggests that the main reason for the lower rate is the underestimated covariance matrix due to insufficient number of clustered descriptors. However, further increasing of the threshold does not improve the results due to significant number of incorrect matches in the clusters. Compared with the state-of-the-art rate of 81.4% in [5], our best result is 75.5% for 20-dim *P* projections. This is due to the fact that in contrast to [5] the spacial location of features has not been used. The recognition rate of 75.5% is 11% higher than 64.4% obtained with original SIFT and 10% higher than the best result of 65.2% from [7].

SIFT(128)	PCA(36)	PCA(30)	PCA(20)	PCA(10)	P(36)	P(30)	P(20)	P(10)
64.4%	64.9%	65.1%	64.9%	65.7%	74.8%	74.7%	75.5%	75.2%

Figure 5. Scene recognition rates of original SIFT feature and projected features(level=4, branch=20). The results are averages of 10-runned experiments.

Conclusion and discussion

We have presented and evaluated methods to estimate linear discriminant projections for dimensionality reduction of local image descriptors. We proposed a method to obtain discriminant projections without ground-truth data in contrast to previously proposed approaches [3, 11]. Experiments show that the proposed strategy not only reduces the feature dimensionality, but also yields better results than the original SIFT descriptor and PCA projected descriptors in various application scenarios. Our methods can be easily applied to any other local image descriptor.

Extensive evaluation indicates different improvements depending on the application. In wide baseline matching where the descriptor variations can be modeled well with rigid

transformations the improvement can be significant. We achieved small but consistent improvement in patch matching. The separation of matched and unmatched features for this data is already high with the original descriptors and there is little scope for improvement. In the case of category recognition where the intra-class variability is high the gain is significant compared to the original SIFT and PCA projected descriptors. In future work, we will explore more complex simulation models including photometric transformations as well as apply the proposed method to improve fast search in large datasets.

Acknowledgment

This research was supported by EU VIDI-Video IST-2-045547 and UK EPSRC EP/F0034 20/1 grants.

References

[1] K. Fukunaga. *Introduction to Statististical Pattern Recognition*. Academic Press, 1990.

[2] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *ICCV* , 2005.

[3] G. Hua, M. Brown, and S. Winder. Discriminant embedding for local image descriptors. In *ICCV* , 2007.

[4] Y. Ke and R. Sukthankar. PCA-SIFT: a more distinctive representation for local image descriptors. In *CVPR*, 2004.

[5] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR* , 2006.

[6] B. Leibe, K. Mikolajczyk and B. Schiele Efficient Clustering and Matching for Object Class Recognition. In *BMVC* , 2006.

[7] Fei-Fei Li and P. Perona. Bayesian Hierarchical Model for Learning Natural Scene Categories. In *CVPR* , 2005.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[9] J. Matas, O. Chum, U. Martin, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, 2002.

[10] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[11] K. Mikolajczyk and J. Matas. Improving descriptors for fast tree matching by optimal linear projection. In *ICCV*, 2007.

[12] J. Robinson. Covariance matrix estimation for appearance-based face image processing. In *BMVC*, 2005.

[13] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. In *SIGGRAPH* , 2006.

[14] S. A. Winder and M. Brown. Learning local image descriptors. In *CVPR*, 2007.