

Learning Linear Discriminant Projections for Dimensionality Reduction of Image Descriptors

Hongping Cai, Krystian Mikolajczyk, *Member, IEEE*, and Jiri Matas *Member, IEEE*,

Abstract—In this paper we present Linear Discriminant Projections (LDP) for reducing dimensionality and improving discriminability of local image descriptors. We place LDP into the context of state-of-the-art discriminant projections and analyze its properties. LDP requires large set of training data with point-to-point correspondence ground truth. We demonstrate that a training data produced by a simulation of image transformations leads to nearly the same results as the real data with correspondence ground truth. This makes it possible to apply LDP as well as other discriminant projection approaches to the problems where the correspondence ground truth is not available such as image categorization. We perform an extensive experimental evaluation on standard datasets in the context of image matching and categorization. We demonstrate that LDP enables significant dimensionality reduction of local descriptors and performance increases in different applications. The results improve upon the state-of-the-art recognition performance with simultaneous dimensionality reduction from 128 to 30.

Index Terms—Linear discriminant projections, dimensionality reduction, image descriptors, image recognition, image matching.

1 INTRODUCTION

MANY recent and successful computer vision approaches are based on local image descriptors. Local descriptors have been applied to image retrieval, recognition, panorama building, robot navigation, visual data mining, text matching, biometrics etc. Significant effort has been made to develop discriminative descriptors such as Geometric Blur [1], SIFT [14], GLOH [15] and the recent DAISY [17], color-SIFT [46]. The descriptors differ in the design and implementation, each trying to optimize the general performance. In this paper we propose a general method for improving the descriptors and reducing their dimensionality by learning their discriminant projections from sample data.

In most application scenarios local descriptors are used to establish correspondences between similar parts of images. The descriptors are characterized by properties such as invariance, robustness, distinctiveness, compactness, and scalability. The descriptors can be made insensitive to small image perturbations, for example, by quantization or integration, that is robust to those perturbations. The level of distinctiveness is related to the entropy of the descriptor. Compactness and distinctiveness are two competing properties and improving the first without decreasing the second would already be a significant progress in scalability. The compactness is directly related to the number of feature dimensions and it is crucial for large scale applications that become the main interest of the research community. These various

properties are related and cannot be optimized at the same time. Existing designs of local descriptors propose different property trade-offs driven by requirements of particular applications or even a dataset only. Furthermore, various interest point detectors select different types of local image patterns that define the support regions for the descriptors but also introduce different type and amount of noise [22]. Discriminant projections can learn a descriptor given a sample of the data, which allows us to avoid re-designing the algorithms or tuning their parameters experimentally. In many computer vision tasks, higher performance has been achieved by increasing the sampling density and descriptor complexity. In contrast, our aim is to improve the performance by projecting local descriptors into more discriminant but fewer dimensions.

In this paper, we build on the top of our initial work from [12], present Linear Discriminant Projections (LDP), analyze in depth their properties as well as relations to other approaches and show that it outperforms PCA. In the context of interest point descriptors very few attempts have been made to use discriminative methods for reducing dimensionality [11], [12], in contrast to widely used PCA [19], [15], [20], [21]. Discriminative methods may lead to significant improvements in a particular setup if trained on annotated sample data from that setup.

LDP requires intra-class covariance, which is estimated on a dataset with point-to-point correspondence ground truth. The ground truth provides *matched* descriptor pairs, where each pair represents similar image pattern or comes from the same physical point of the scene/object. However, the requirement for annotated training data makes this method less attractive than PCA. While in wide baseline matching it is possible to establish unique correspondences by applying geometric

- H. Cai is with School of Electronic Science and Engineering, National University of Defense Technology, China. H. Cai and K. Mikolajczyk are with the Centre for Vision, Speech and Signal Processing, University of Surrey, UK. E-mail: hongpingcai@hotmail.com, k.mikolajczyk@surrey.ac.uk.
- J. Matas is with Faculty of Electrical Engineering, Czech Technical University, Czech Republic. Email: matas@cmp.felk.cvut.cz

constraints, it is more ambiguous to define such correspondences in object recognition and it requires tedious manual annotation. Global image transformations such as homography cannot be used here as similar features may occur in different location on objects of the same category (cf. section 4.1, Table 2). Similarly, object or scene parts of the same semantic meaning may have very different appearance, thus different descriptors. Geometric blur [1] or other techniques based on similar idea [23], [18], [17], model the signal variations by averaging it over a range of acceptable geometric transformations. However, this idea has not been used for learning discriminant projections of local descriptors. In this paper, we show that the training based on simulated data can be successfully used for obtaining discriminant projections, and that it overcomes the issues related to data annotation and makes the LDP approach applicable to any dataset and any descriptor. Our simulation strategy can also be easily used for any supervised dimensionality reduction techniques which require matched features for training. This opens new application possibilities for such techniques. We evaluate the discriminant projections using new training approach and different datasets. The method improves state-of-the-art SIFT descriptor and image categorization approach from [20], [25], with the number of descriptor dimensions reduced from 128 to 30.

LDP has a number of appealing properties that make it a method of choice for projecting descriptors. LDP is a global technique and it is not as sensitive to noise as local methods discussed in section 2. Unlike in local methods, there is no need to search for the k -nearest neighbors for each point, which makes it faster and independent of k NN parameters. In contrast to many distance metric learning approaches, its complexity is low. LDP projections can be directly applied to the descriptors and there is no need to re-optimize the projections if the required number of output dimensions changes. Finally, LDP has been reported to improve matching performance in [11], to result in more efficient search in various tree-like data structures in [12] and to increase image categorization accuracy in [13]. Consistent improvements in different test scenarios indicate that the approach generalizes well for different vision problems.

The main contributions of the paper are:

- 1) Linear discriminant projections applied to local image descriptors and their formulation in a graph embedding framework [2].
- 2) In depth analysis of LDP as well as a review of other discriminant projection methods.
- 3) An approach to train the projections for any data by generating simulated training set also applicable to other discriminant projection methods.
- 4) Extensive evaluation on 2 image matching and 3 image categorization benchmarks.
- 5) State of the art results in all 3 recognition benchmarks with significant dimensionality reduction.
- 6) Projection matrices and software libraries for re-

ducing dimensionality with discriminant projections made available to the community for comparisons [31].

This article is structured as follows. We first review the related literatures on the projection methods potentially applicable to local features in section 2. In section 3, we analyze linear discriminant projections. A simulating strategy for learning the projection vectors from any unannotated data is proposed in section 4. Next, experiments are carried out for image matching in section 5 and for recognition in section 6. Finally, conclusions and discussion on this work end this article.

2 RELATED WORK

Dimensionality reduction techniques can be categorized into unsupervised and supervised methods that use data without and with ground truth, respectively. The most widely used linear approach that belongs to the first category is PCA. Unsupervised nonlinear techniques include manifold learning approaches such as ISOMAP [3], Locally Embedded Analysis [4], Laplacian Eigenmap [5], which have been reviewed in [35]. Despite the fact that the nonlinear methods achieve considerable performance improvements on some datasets, their crucial limitation is that the embedding does not generalize well from training to test data [35]. The unsupervised approaches are convenient to use but their discriminating capabilities are limited since no class discriminatory information is used during training. We therefore focus on supervised LDP and discuss the related supervised linear approaches.

Supervised approaches aim to map the original space to a lower dimensional space and to preserve the class discriminatory information from labeled point. Although many techniques with different optimization objectives have been proposed, they share the same goal. The distance between the points with the same label is reduced in the projected space while at the same time differently labeled points are made apart to avoid the problem of shrinking the entire data space. The differently labeled points can be either from different classes (any or nearest neighbors) or any points regardless the label. The method is global [11], [12], [7], [26], [34] if all points with the same label are considered or local [8], [2], [33], [9] if only k nearest neighbor pairs are used. Below, we discuss a number of methods that focus on the same goal. The projections provided by these methods are illustrated in Fig. 1.

All dimensionality reduction methods can be formalized by the following notation that is used in the remainder of the paper. Given a set of data points $\mathbf{x}_1, \dots, \mathbf{x}_N$, $\mathbf{x}_i \in \mathbb{R}^m$, the goal of dimensionality reduction is to find a mapping $F : \mathbb{R}^m \mapsto \mathbb{R}^{m'}$, $m' < m$: $\mathbf{y}_i = F(\mathbf{x}_i)$, $\mathbf{y}_i \in \mathbb{R}^{m'}$. If the mapping is linear, it can be formulated as $\mathbf{y}_i = W^T \mathbf{x}_i$ by considering $W \in \mathbb{R}^{m \times m'}$ as a linear projection matrix $W = [\mathbf{w}_1, \dots, \mathbf{w}_{m'}]$.

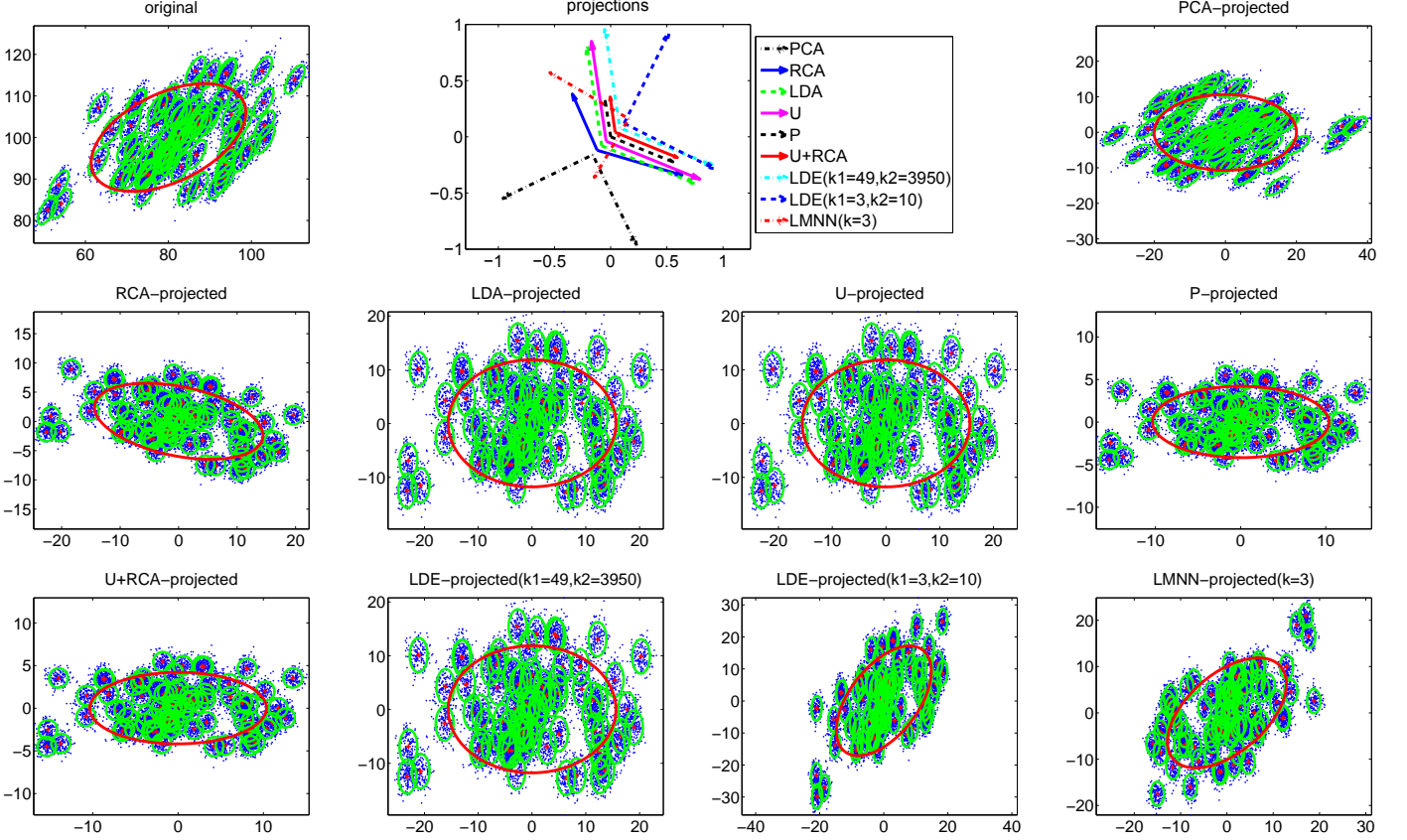


Fig. 1. 2-dim simulated feature set of 80 Gaussian distributed clusters and their projections. Each cluster has 50 features generated with the same covariance. All the features in the same cluster are considered matched, while those from different clusters are unmatched. LDP projections P and U have the same orientation, but different magnitude. A combination of RCA and U -projected features is equivalent to P -projected features. LDA produces the similar orientations as U in this set. LDE finds the same projections as LDP if it considers all matched features (49) and all unmatched features (3950). LDE projections differ from LDP, if nearest neighbors ($k = 3$) are considered only.

LDA. Among supervised approaches, the most popular method is LDA, with the objective to maximize the ratio of between-class (S_B) to within-class (S_W) scatters along direction \mathbf{w} :

$$\mathbf{w}_{LDA} = \arg \max_{\mathbf{w}} \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \quad (1)$$

Where $S_W = \sum_{c=1}^{N_c} \sum_{i:c_i=c_j} (\mathbf{x}_i - \bar{\mathbf{x}}^c)(\mathbf{x}_j - \bar{\mathbf{x}}^c)^T$, $S_B = \sum_{c=1}^{N_c} n_c (\bar{\mathbf{x}}^c - \bar{\mathbf{x}})(\bar{\mathbf{x}}^c - \bar{\mathbf{x}})^T$, $\bar{\mathbf{x}}^c$ and $\bar{\mathbf{x}}$ denote the mean vectors of the c -th class and the whole dataset, respectively. n_c is the number of samples in the c -th class and N_c is the number of classes. S_W captures the data distribution in each class, and S_B represents the separation of the class means. LDA is optimal in the Bayes sense, if all classes have identical Gaussian distribution, which is not always the case. LDA can have at most $N_c - 1$ projections and suffers from small sample size problem in the case of high-dimensional data, which may lead to singular within-class scatter matrix.

LDE. Linear Discriminant Embedding (LDE) [8] was proposed to integrate the information of nearest neighbors and class relations between data points. The same

method called Marginal Fisher Analysis (MFA) was proposed in [2] within their graph embedding framework. The idea is to maintain the original neighbor relations of points from the same class while pushing apart the neighboring points of different classes. Compared to LDA, LDE has two merits. First, the number of projections are not limited. Second, there is no assumption about the Gaussian distribution.

LDP. An embedding technique was proposed in [11], which can be viewed as a global version of LDE or MFA. Based on the same idea, our LDP approach initially proposed in [12], simultaneously diagonalizes the inter- and intra-class covariance matrices. We discuss the properties of this approach in more detail in section 3. This global dimensionality reduction method can also be included in the graph embedding framework from [2] as we demonstrate in section 3.2.

MDML. Linear projections have close relation with Mahalanobis Distance Metric Learning (MDML) [35], where a positive semi-definite matrix $M \in \mathbb{R}^{m \times m}$ is learned to define a new distance metric: $d_M^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T M (\mathbf{x}_i - \mathbf{x}_j)$. M can be decomposed as $M = V^T V$,

where $V = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{m'}]$ corresponds to a linear projection matrix from the original space to a new space. If M is a full-rank matrix, then $m' = m$ and the projected space has the same dimensionality as the original space. However, many MDML methods combine metric learning with dimensionality reduction by solving for a rectangular projection matrix $V \in \mathbb{R}^{m \times m'}$ ($m' < m$). Below, we discuss the commonly used supervised MDML approaches such as GDML, RCA, NCA, SVM-RC and LMNN.

GDML, RCA and NCA. The goal of Global Distance Metric Learning method [7] is to minimize the sum of all distances between same-labeled points under the constraint that the sum of distances between different-labeled points is large. Relevant Component Analysis (RCA) [26] learns a global linear transformation from the equivalence constraints and applies whitening transform to the intra-class covariance. Neighborhood Component Analysis (NCA) [33] is a local method that extends nearest neighbor classifier with metric learning. It learns the projections by maximizing k-nearest neighbor classification with leave-one-out cross validation. However, it is argued in [35] that NCA tend to overfit to training data in high dimensional spaces and it suffers from the convergence as well as scalability problems.

SVM-RC and LMNN. In supervised distance metric learning, SVM-like convex optimization for learning from relative comparisons (SVM-RC) [34] has also been investigated. The constraints involve relative comparisons of individual image pairs in contrast to the approaches discussed above with only one constraint on the global sum of distances. Two typical approaches SVM-RC [34] and Large Margin Nearest Neighbor (LMNN) [9] have achieved considerable performance in some recognition benchmarks. Both methods share the same constraints except that all the same labeled pairs are considered in SVM-RC while only k -nearest neighbors of the same labeled points are used in LMNN. Furthermore, the objective function of SVM-RC minimizes the hinge loss with the ℓ_2 regularization, while LMNN minimizes the distance of the same labeled image pairs and the hinge loss. The motivation for these two methods is very similar to the graph embedded approaches [2] although SVM-RC and LMNN cannot be formulated in this framework. SVM-RC seeks distance metric to bring all the same labeled points closer and keep the differently labeled points apart, which is very similar to the idea of LDA, LDP [11], [12] and GDML[7]. LMNN [9] shrinks the k-nearest neighbors of the same class while separating all different class pairs, which is similar to LDE [8] / MFA [2]. This similarity is illustrated in Fig. 1 for $k = 3$.

Table 1 summarizes the optimization objectives of the discussed approaches. \mathcal{S} is a set of pairs of points that belong to the same class and \mathcal{D} is a set of pairs of points that belong to different classes. The method is global or local depending whether it uses all pairs from set \mathcal{S} or kNN pairs only. Local methods can also use the

heat kernel on the distance between points to weight the influence of kNN pairs.

TABLE 1

Linear projections. \mathcal{S} and \mathcal{D} stand for the same labeled points and differently labeled points, respectively. 'all' and 'kNN' indicate whether all the pairs (global) are considered or k-nearest neighbors only (local).

dataset \ method	reduce distance		increase distance			
	\mathcal{S}		\mathcal{D}		$\mathcal{S} \cup \mathcal{D}$	
	all	kNN	all	kNN	all	kNN
PCA					✓	
LDA	✓				✓	
LDE [8], [2]		✓		✓		
LDP [11], [12]	✓		✓			
GDML [7]	✓		✓			
NCA [33]		✓				✓
RCA [26]	✓				✓	
SVM-RC [34]	✓				✓	
LMNN [9]		✓	✓			

3 LINEAR DISCRIMINANT PROJECTIONS

In this section, we present details on linear discriminant projections, demonstrate equivalence of the methods proposed in [12] and [11], as well as discuss the relations to other methods. In [11], a projective direction \mathbf{u} is designed to maximize the ratio of variance of differently labeled points (\mathcal{D}) to that of same labeled points (\mathcal{S}). It can be formulated as follows:

$$\begin{aligned} \mathbf{u}_{\text{LDP}} &= \arg \max_{\mathbf{u}} \frac{\sum_{(i,j) \in \mathcal{D}} \|\mathbf{u}^T \mathbf{x}_i - \mathbf{u}^T \mathbf{x}_j\|^2}{\sum_{(i,j) \in \mathcal{S}} \|\mathbf{u}^T \mathbf{x}_i - \mathbf{u}^T \mathbf{x}_j\|^2} \\ &= \arg \max_{\mathbf{u}} \frac{\mathbf{u}^T C_{\mathcal{D}} \mathbf{u}}{\mathbf{u}^T C_{\mathcal{S}} \mathbf{u}} \end{aligned} \quad (2)$$

Where $C_{\mathcal{D}}$ and $C_{\mathcal{S}}$ represent the inter- and intra-class covariance matrices of differently labeled points (unmatched features in image descriptor space) and same labeled points (matched features), respectively.

$$C_{\mathcal{D}} \stackrel{\text{def}}{=} \sum_{(i,j) \in \mathcal{D}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \quad (3)$$

$$C_{\mathcal{S}} \stackrel{\text{def}}{=} \sum_{(i,j) \in \mathcal{S}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \quad (4)$$

Note that these are not the same matrices as the between-class S_B and within-class scatters S_W in equation (1) for LDA, although they are related (see section 3.3). The solution is the generalized eigenvectors:

$$U = \text{eig}(C_{\mathcal{S}}^{-1} C_{\mathcal{D}}) \quad (5)$$

The projection matrix is $U \in \mathbb{R}^{m \times m'}$, with $m' \leq m$ eigenvectors corresponding to the m' largest eigenvalues.

LDP was initially proposed in [12] and consists of two parts. The first one is the inverse of the square root of intra-class covariance matrix $C_{\mathcal{S}}^{-\frac{1}{2}}$, which is used for whitening of the original feature space. The second

part is PCA of the inter-class covariance matrix in the whitened space $\tilde{Y} = \{C_S^{-\frac{1}{2}}\mathbf{x} | \mathbf{x} \in X\}$:

$$P = C_S^{-\frac{1}{2}} \cdot \text{eig}(C_S^{-\frac{1}{2}} C_D C_S^{-\frac{1}{2}}) \quad (6)$$

After the projection the features are normalized such that

$$\|P^T \mathbf{x}\| = \|U^T \mathbf{x}\| = c \quad (7)$$

where c is a constant. Projection matrix $U \in \mathbb{R}^{m \times m'}$ and $P \in \mathbb{R}^{m \times m'}$ are used to refer to the two linear discriminant projections from [11] and [12], respectively.

3.1 Relations Between LDP and U

Although the approaches to obtain the projection vectors LDP (P) [12] and U [11] differ, both methods use the same covariance matrices of pairwise matched feature distances and pairwise unmatched feature distances. Furthermore, these methods can be considered equivalent as they find the same projection directions. The analytical proof is in Appendix A and the experimental results in section 5 validate this equivalence. Fig. 1 shows P and U projections for a simulated feature set. Fig. 2 shows P and U for Normalized Gray value (NG) patches, i.e., gray values normalized to zero-mean and one-variance, as well as for SIFT features. Both figures show that the projection directions are the same.

However, the magnitudes of the projective vectors differ, in particular the sign difference can be observed in Fig. 2. According to equation (6), P rotates the feature space as well as scales the dimensions by $C_S^{-\frac{1}{2}}$, while U only rotates the space. If P_i and U_i define the i -th projective vectors of P and U , then $\|P_i\| \neq 1$ and $\|U_i\| = 1$. By comparing the U - and P -projected features in Fig. 1, we observe that both projections have diagonalized the intra- and inter-class covariances and P additionally applies the whitening to the intra-class covariance. Since RCA is such a whitening transform, a combination of U and RCA is equivalent to P , which is verified in Fig. 1.

Descriptors projected with P and U should lead to similar performance due to the same projection directions. The whitening process however normalizes the space and makes it optimal for nearest-neighbor search with Euclidean distance, thus P projections may provide better results in applications relying on NNs. Comparison results are presented in section 5.2.

In Fig. 2, we also observe that projections P and U are concentrated on the center of the patches in contrast to PCA. Note that similar spatial weighting is also implemented in SIFT using Gaussian kernel.

3.2 Graph Embedding of LDP

A number of methods for dimensionality reduction have recently been included in a general framework [2] called Graph Embedding. It unifies dimensionality reduction techniques including Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) [27],

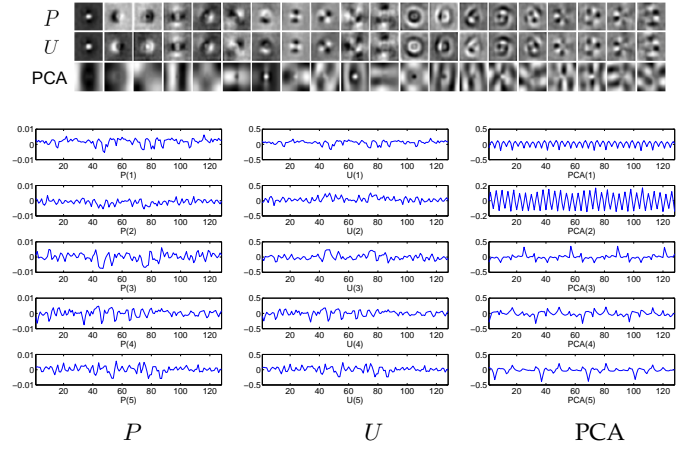


Fig. 2. Projective vectors of P , U and PCA. 100,000 patch pairs (50% matched and 50% unmatched) randomly picked in Liberty set from [10] are used for estimating the projections. Top: The top 20 projections from left to right of Normalized Gray value (NG) feature. The patches have been normalized to zero-mean one-variance. Bottom: The top 5 projections of SIFT feature.

ISOMAP [3], Locally Embedded Analysis (LEA) [4], Laplacian Eigenmap (LE) [5], Locality Preserving Projection (LPP) [6], and Marginal Fisher Analysis (MFA) [2]. We extend this set by formulating Linear Discriminant Projections (LDP) in this framework.

Following [2], an intrinsic graph $G = \{X, W\}$ and a penalty graph $G' = \{X, W'\}$ are two undirected weighted graphs with vertex set $X = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$, and with the affinity matrices $W \in \mathbb{R}^{N \times N}$ and $W' \in \mathbb{R}^{N \times N}$. Their elements W_{ij} and W'_{ij} represent the edge weights between features \mathbf{x}_i and \mathbf{x}_j . It has been shown in [2] that many linear dimensionality reduction approaches share the same graph-preserving criterion.

$$\begin{aligned} \mathbf{w}^* &= \arg \min_{\mathbf{w}^T X B X^T \mathbf{w} = d} \sum_{i \neq j} \|\mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{x}_j\|^2 W_{ij} \\ &= \arg \min_{\mathbf{w}^T X B X^T \mathbf{w} = d} \mathbf{w}^T X (D - W) X^T \mathbf{w} \end{aligned} \quad (8)$$

where D is a diagonal matrix with diagonal elements $D_{ii} = \sum_j W_{ij}$, d is a constant and B is a constraint matrix defined to avoid a trivial solution of the objective function. PCA, LDA, LDE/MFA, LPP, LEA can all be included in this framework with different definition of W and B in (8). We formulate LDP in the Graph Embedding framework by modifying equation (2):

$$\begin{aligned} \mathbf{u}_{LDP} &= \arg \max_{\mathbf{u}} \frac{\mathbf{u}^T X (D' - W') X^T \mathbf{u}}{\mathbf{u}^T X (D - W) X^T \mathbf{u}} \\ &= \arg \min_{\mathbf{u}^T X (D' - W') X^T \mathbf{u} = d} \mathbf{u}^T X (D - W) X^T \mathbf{u} \end{aligned} \quad (9)$$

where $W_{ij} = 1$ if $(i, j) \in \mathcal{S}$, $W'_{ij} = 1$ if $(i, j) \in \mathcal{D}$, and $B = D' - W'$. The formula is the same for LDE/MFA [2] except that W and W' are defined there locally i.e., $W_{ij} = 1$, if $(i, j) \in \mathcal{S}_{kNN}$ and $W'_{ij} = 1$, if $(i, j) \in \mathcal{D}_{kNN}$. Thus,

LDP can be viewed as a global version of LDE/MFA and it is not as sensitive to noise as local methods that use nearest neighbors only. Another advantage over LDE is that LDP can be trained significantly faster as there is no need to search for the k -nearest neighbors for each point. Moreover, the performance of LDE depends on parameters k_1 for S_{kNN} and k_2 for \mathcal{D}_{kNN} , that have to be chosen experimentally.

3.3 Relations Between LDP and LDA

It has also been demonstrated in [2] that LDA can be formulated in the graph embedding scheme, with $W_{ij} = \delta_{c_i, c_j} / n_{c_i}$, $W'_{ij} = 1/N$. Hence, we observe that LDP and LDA have the same W when each class contains the same number of samples $n_{c_i} = n_{c_j}$. LDA inter-class covariance is typically estimated from matched and unmatched pairs ($\mathcal{S} \cup \mathcal{D}$), while LDP uses unmatched pairs (\mathcal{D}) only. If the number of data points is large compared to the number of classes then $C_{\mathcal{D}} \approx C_{\mathcal{S} \cup \mathcal{D}}$ since $|\mathcal{S}| \ll |\mathcal{D}|$. Both methods assume Gaussian distribution of data points. The main difference between LDA and LDP is therefore in the estimation of the covariance matrices which in LDP are based on pair-wise descriptor differences and in LDA on point-to-mean differences.

The illustration in Fig. 1 shows that for this particular experimental setup, LDA and LDP result in similar projections. There are 80 clusters and each cluster contains same number of features (50), thus $|C_{\mathcal{S}}| = 1.24\%|C_{\mathcal{D}}|$, $|C_{\mathcal{S}}| = \binom{50}{2} \times 80$, $|C_{\mathcal{D}}| = (79 \times 50) \times (80 \times 50)/2$.

4 COVARIANCE ESTIMATION

The main problem in computing discriminant projections is the estimation of the intra-class covariance matrix $C_{\mathcal{S}}$. In this section we investigate methods for estimating this matrix and discuss the impact they may have on the projection vectors. The main interest is in the simulation approach which does not require annotated training data.



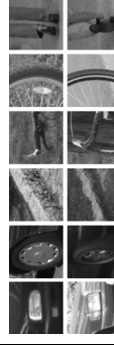
4.1 Feature Transformations

The intra-class covariance matrix estimation requires matched features and inter-class matrix uses unmatched features. The definition of matched and unmatched features may differ to some extent in different applications. Three categories of matched pairs are illustrated in Table 2. In matching images of planar scenes or 3D structures, matched features are local descriptors from the same physical point on the object, while in image or object categorization, matched features correspond to points on the objects that are visually similar and usually belong to the same object category. The main sources of descriptor variations and noise in local feature applications are:

- geometric image transformations, due to viewpoint change and camera zoom,

TABLE 2

Matched features in various vision problems. From planar scenes to image/object categorization, the appearance changes of matched features increase.

planar scene	3D scene	image category
same physical points of the scene		corresponding object/scene parts
		
		appearance variations
		occlusion
		geometric and photometric transformations, noise, spatial, brightness and color quantization

- photometric transformations, due to brightness and color of scene illumination or varying camera exposure,
- occlusion and self occlusion of 3D structures due to viewpoint change or moving objects,
- noise introduced by inaccuracy of the region detector which mainly consists of small geometric deformations,
- appearance/shape variations of object/scene parts that belong to the same category (in image categorization).

It is possible to model analytically to some extent the variations from the first two variations, if the patch is planar, unfortunately the other variations are unpredictable and can only be captured from training examples. The computation of the projections should therefore be based on the training examples that represent the statistics of the data as well as statistics of the region detector. In section 5.2, we investigate to what extent the variations between matched features illustrated in Table 2 can be modeled by simulation.

4.2 Ground-Truth Data

In [11] and [12], the linear discriminant projections require a large amount of ground-truth data for intra-class covariance estimation. Matched feature pairs are produced by bundle adjustment in images of 3D scenes [10] and by homography of image pairs [31]. The matched pairs are then used to compute the intra-class covariance according to equation (4). To overcome the problem of insufficient training data especially for high dimensional features, this estimation is often followed by a regularization [11].

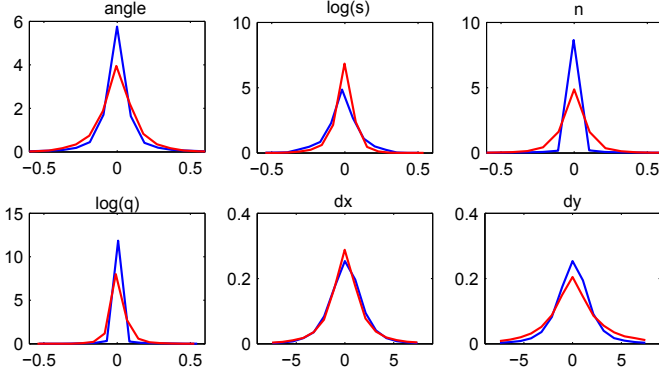


Fig. 3. Distribution of 6 affine parameters (rotation, scale $\log(s)$, skewing n , stretching $\log(q)$, dx translation and dy translation) in affine transform between matched regions from 15 image sequences [31]. Two regions are considered matched if their overlap error is less than 50%. Blue and red curves represent distribution for multi-scale Harris and MSER detectors, respectively.

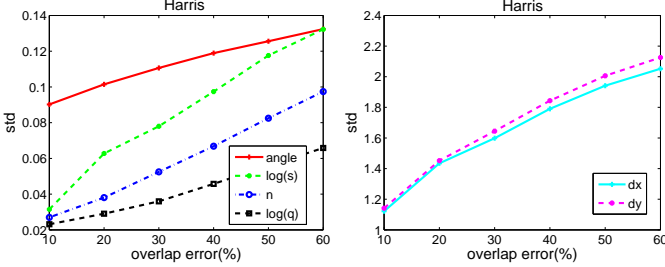


Fig. 4. Standard deviations of 6 affine parameters for matched region pairs of multi-scale Harris as a function of overlap error between matched regions.

4.3 Simulated Data

In order to learn the projections for data without the ground truth, we investigate possibilities of modeling the expected variations artificially. It has been observed in other experiments [22], [15] that the photometric invariance of features is sufficient to survive the most common illumination changes, therefore there is little benefit from explicit modeling of such variations. Furthermore, we assume uniform noise distribution from occlusion as well as quantization and focus on geometric transformations in a similar way to [1], [23], [18], [17]. Given that the descriptors are computed locally and assuming local smoothness, affine transformation is sufficient to model the geometric changes. The affine transformation can be decomposed to rotation, scaling, skewing, stretching and translation. The remaining issue is therefore the estimation of the parameter values that should be used to model the affine changes. To address this problem we investigate the geometric changes between matched patches in real data with ground truth.

4.3.1 Parameter Estimation

We use 15 image sequences from [31] with homography ground truth for estimating the distribution of affine

parameters. The matched regions are identified by their overlap error using the homography. These regions are then normalized with their scales and dominant angles [14]. The transformation parameters between normalized regions are estimated using the homography. These parameters can be considered as errors remaining after the normalization or as the inaccuracy of the interest point detector.

Fig. 3 shows the distributions of the 6 parameters for multi-scale Harris (blue) and MSER (red) detectors, which can be approximated by normal distributions. We also observe that these distributions differ for multi-scale Harris and MSER detectors. This confirms the observations from [15] that different parameter estimation accuracy can be achieved for different types of regions provided by various detectors. For example, the rotation angle seems to be more accurate for multi-scale Harris than MSER. This also holds for skewing n and stretching $\log(q)$, while scale $\log(s)$ estimation error is smaller for MSER. MSER features are centered on blob-like structures while Harris on strong gradient. Scale estimation is known to be more reliable for blob like structures than for corners [16] but the dominant gradient orientation is more ambiguous for blobs which explains the observations. Fig. 4 shows how the overlap error affects the variance for each transform. Intuitively, the larger the overlap error is, the more variation in the parameter values, which is consistent for all parameters in Fig. 4.

4.3.2 Discussion

Our simulation strategy for estimating the projections can also be considered as modeling variation in descriptors by simulating the image transformations to improve its robustness to small signal changes. This idea is similar to the geometric blur [1], which averages the signal over a range of acceptable transformations (small affine transformations are also considered in [1], [23], [18], [17]). This average is computed by convolving the signal with a spatially varying kernel. If $I(x)$ and $I(T(x))$ are matched patches and T is a transformation from a space of transformations \mathcal{T} , then the aim of the geometric blur is to integrate a patch over all possible $T \in \mathcal{T}$. We extend this concept to a descriptor where $\phi(I)$ is an operator for computing a descriptor vector (instead of $\phi(I(x))$ to simplify the notation). The geometric blur on patches (GB_I) and descriptors (GB_ϕ) is defined as:

$$GB_I = \int_{\mathcal{T}} I(T(x))p(T)dT \quad (10)$$

$$GB_\phi = \int_{\mathcal{T}} \phi(T(I))p(T)dT \quad (11)$$

The intra-class covariance integrates over all possible transformations $T \in \mathcal{T}$ and image patches $I \in \mathcal{I}$:

$$E\{\phi(I) - \phi(T(I))\}\{\phi(I) - \phi(T(I))\}^T = \int_{\mathcal{I}} \phi(I)\phi^T(I)p(I)dI - 2 \int_{\mathcal{I}} \phi(I)GB_\phi p(I)dI + \int_{\mathcal{I}} GB_\phi^2 p(I)dI$$

The densities $p(T)$ and $p(I)$ model the transform and image probability, respectively. The terms GB_ϕ and GB_{ϕ^2} are then geometric blurs on the descriptors. While it is possible to model the integral of the signal transformations by a convolution with an appropriate kernel $GB_I(x) = \int_y I(x-y)K_x(y)dy$, as in [1], GB_ϕ cannot be directly implemented as a convolution with the descriptor. Instead, the transformations can be applied to the image region before computing the descriptors, which can then be used to learn a robust feature.

5 MATCHING EXPERIMENTS

In this section we demonstrate the performance of the proposed method in the context of wide baseline matching. We demonstrate that LDP can adapt a descriptor to this task and produce better results than other approaches.

5.1 Setup

5.1.1 Datasets

We use two benchmark datasets from [10] and [31] for image matching. Dataset from [10] consists of clusters of patches sampled from 3D reconstructions of the Statue of Liberty, Notre Dame and Half Dome (Yosemite). Each of the sets contains two subsets of patches detected with DoG [14] and multi-scale Harris [22] detectors. Each subset contains 400k patches with patch-to-patch correspondence ground truth.

We also perform experiments on image sequences from [31] with homography ground truth. We use 15 sequences with 198 image pairs for training and 3 sequences with 22 image pairs for testing, which include rotation, scale, rotation-scale and viewpoint changes. We refer to this data by Oxford. Patches from this set are considered matched if the overlap error between the corresponding regions is less than 50%. The data subsets and the number of matched (S) and unmatched (D) patch pairs used for training and testing in different experiments are summarized in Table 3.

5.1.2 Descriptors

We use grayvalue patches (NG) normalized to 0-mean and 1-variance and SIFT as the baseline descriptors. The patches from [10] are 64×64 pixels, with a canonical scale and orientation but all the experiments are performed on the center 36×36 pixels part to avoid boundary problems in the simulation process. We refer to the descriptors as DoG-NG, DoG-SIFT and Har-SIFT.

In addition, for the dataset from Oxford [31], we extract patches with multi-scale Harris [22] as well as MSER [28] and compute SIFT descriptors (Har-SIFT, MSER-SIFT).

The projections are trained and applied to the original descriptors. The projected feature vectors are normalized to a constant length [11]. This normalization stabilizes the descriptors and improves their performance.

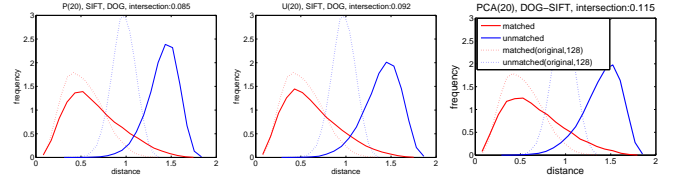


Fig. 5. *Experiment M01*. Distance distributions of matched and unmatched features for the 1296-dim NG features and their 15-dim projections. Left: P . Middle: U . Right: PCA. The intersection/union area ratio for the projected feature is given at the top of the figures and can be compared to the original one (0.281).

5.1.3 Evaluation

We follow the evaluation protocols that have been introduced with the datasets. The Euclidean distance between each pair of test patches is calculated and thresholded. The matches are then compared with the ground truth. ROC curve is obtained for the data from [10] by varying the distance threshold. The true positive rate is given by TP/S and false positive rate is FP/D , where TP is the number of correctly matched pairs, FP is the number of incorrectly matched pairs, S is the total number of matched test feature pairs and D is the number of unmatched feature pairs used for the evaluation. The performance is then reported by the true positive score (TP/S) at equal error rate (eer matching score).

Precision-recall curves are estimated for data from Oxford [31], where the recall is TP/S and the precision is $TP/(TP + FP)$. The area of the precision-recall curve (average precision) is used for evaluation.

5.2 Matching Results

5.2.1 Experiment M01. Distance Distributions

This experiment compares the distance distributions for matched and unmatched NG features before and after the projections as shown in Fig. 5. The ratio of the intersection area between the distance distributions of matched and unmatched features to their union area indicates the discriminability of the feature. The intersection area ratio significantly decreases after the projections resulting in a better separation of matched and unmatched features with 15 dimension only. In particular, 15-dim P (0.143) and U (0.148) have lower intersection area than both 15-dim PCA (0.172) and the 1296-dim NG (0.281).

5.2.2 Experiment M02. Performance vs. Dimensionality

Fig. 6 compares the matching performance of low dimensional projected features with that of the original NG on Liberty set. P and U yield best matching score followed by PCA. Compared with PCA and 1296-dim NG, 15 dimensional projections P improve the score by 1.7% and 8.5%, respectively. Consistent improvement are achieved with 1.6% over PCA and 8.4% over NG on Notre-Dame, as well as 1.4% over PCA and 8.1%

TABLE 3
Summary of matching experiments. S and D indicate matched and unmatched feature numbers.

Experiments	data		descriptor
	train	test	
Experiment M01. Fig.5 intra- vs. inter-class distance distributions	Liberty	Liberty (50000 S , 50000 D)	DoG-NG
Experiment M02. Fig.6 performance vs. dimensionality	(50000 S , 50000 D)		DoG-SIFT
Experiment M03. Fig.7 simulation with various transformations	Liberty (50000 S , 50000 D)		
Experiment M04. Fig.8 simulation vs. ground truth	Oxford (~50000 S , ~50000 D)	Oxford (~8000 S , ~50000 D)	Har-SIFT
Experiment M05. Fig.9 simulation vs. ground truth distribution	Liberty	Liberty (50000 S , 50000 D)	MSER-SIFT
			Har-SIFT DoG-SIFT
	Notre-dame	Liberty (50000 S , 50000 D)	DoG-SIFT
	Yosemite		DoG-SIFT
Experiment M06. Fig.10 generalization across detectors and data	Oxford	Oxford (~8000 S , ~50000 D)	Har-SIFT
	Liberty		Har-SIFT DoG-SIFT
Experiment M07. Table 4 simulation vs. ground truth			

over NG on Yosemite. The improvement is larger for low false positive rate as displayed in Fig. 6 (Right). PCA already improves the performance significantly by removing many noisy dimensions. We observe a decreasing performance with increasing dimensionality of NG descriptors, in particular for P -projected features. P -projections scale the dimensions by the eigenvalues of the intra-class covariance which are unstable. This is due to insufficient training data in high dimensional feature space such as NG, but it has little effect on 128-dim SIFT, as shown in Fig. 8.

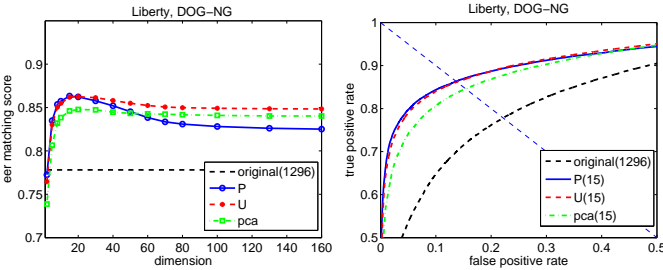


Fig. 6. Experiment M02. Comparison of the matching performance of P , U and PCA with DoG-NG. Left: Matching score at equal error rate (eer matching score) as a function of projected feature dimension. Right: ROC curve for matching with 15-dim projected features compared with the 1296-dim NG. Matching score at eer corresponds to the intersection point of the ROC curve and the diagonal.

5.2.3 Experiment M03. Affine Transformations

In this experiment, we investigate how different combinations of affine changes affect the matching performance. Unlike in the experiment with ground-truth data, no training set is needed, that is, the simulation is directly performed on patches in the test set, which is possible in a practical scenario.

The main issue is the choice of the affine parameters to match the distribution of the real transformations. We first estimate the 6 affine parameter distributions in the patch dataset using an accurate registration software. We remove the large values which are due to the registration error and use the remaining ones to estimate the distributions. Similar to the distributions from

Fig. 3, all 6 parameters are Gaussian distributed with standard deviations $\{\sigma_{\theta_0}, \sigma_{\log(s_0)}, \sigma_{n_0}, \sigma_{\log(q_0)}, \sigma_{x_0}, \sigma_{y_0}\} = \{0.164, 0.120, 0.184, 0.100, 4.81, 4.88\}$. The original patch is transformed with parameters sampled from the Gaussian distributions. We use these 6 standard deviations to initialize the search for optimal patch transformations. We vary the parameter values and test different combinations of affine transformations. For each transform, 6 standard deviations of the Gaussian are tested: $\sigma_* = \frac{i}{5}\sigma_{*0}$ where $i = 0, 1, 2, \dots, 5$ indicates how much variance is allowed for each parameter *. No transformation is done for $i = 0$. As there is little difference between standard deviations for x and y translation, only 5 parameters are considered. Hence, there are 7666 ($=6^5 - 1$) possible combinations.

Fig. 7 displays the results for the best combinations of the individual transformations. Rotation ('20000') yields training patches that gives significantly better projections than any other individual transform. Its performance is even higher than that of the original SIFT descriptor. Among the combinations of 2 changes, 'rotation+scale' produces the highest performance. The performance slightly increases with more changes. The best performance is produced by the combination '45115', which is very close to that obtained by using the training data with the ground truth and higher than the PCA-projected as well as the original SIFT feature.

5.2.4 Experiment M04. Simulation Performance

Fig. 8 compares the performance of projections obtained from simulation parameters '45115' with that of the ground-truth data. The main observation is that the performance of projections from simulated data is very close to the one from the ground-truth data. Both P and U projections of SIFT improve upon the original descriptor as well as its PCA projections for any number of dimensions larger than 10. Moreover, both projections result in very similar improvement. The performance increase is up to 4% with dimensionality reduction from 128 to 20. Interestingly, PCA also improves upon the original SIFT. This is due to the additional normalization (cf. equation (7)) of feature vectors that is performed after the projections.

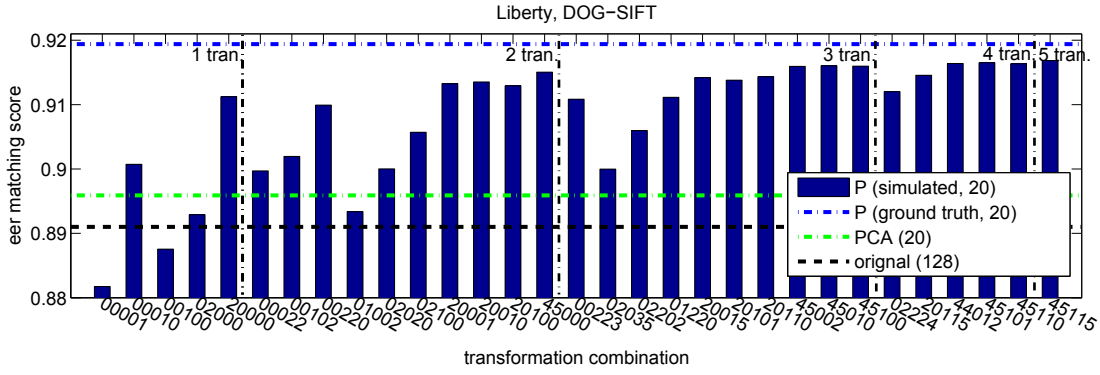


Fig. 7. *Experiment M03*. The matching performance with P -projected 20-dim features from simulated data with various combinations of 5 transformations (rotation, scale, skewing, stretching and translation). The five-digit number indicates the best values for a combination of these transformation parameters i.e. '45100' is the combination of the first three changes, without stretching and translation. Among all different parameters, the best parameters for rotation-scale-skewing combination are $\sigma_\theta = 4/5 \times \sigma_{\theta_0}$, $\sigma_{\log(s)} = 5/5 \times \sigma_{\log(s_0)}$ and $\sigma_n = 1/5 \times \sigma_{n_0}$.

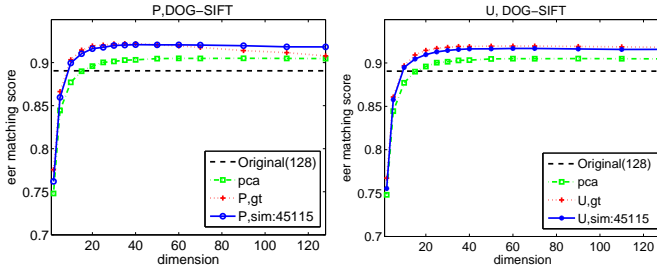


Fig. 8. *Experiment M04*. Comparison of the projection performance obtained from the simulated data (sim) with that from the ground-truth data (gt). The simulated data are generated by the best transformation combination '45115' in Fig. 7.

5.2.5 Experiment M05. Simulation Distance Distributions

Fig. 9 compares the distance distributions of the features generated from the ground-truth data with those generated from the simulated data. Fig. 9 (Left) shows distributions for 20 dimensional P -projected features on Liberty set. The intersection/union area ratio given at the top of the figures is very similar for the ground-truth data and the simulated data. These, as well as the results from *M04*, demonstrate that the simulation can model the real data distribution sufficiently close such that the projections achieve comparable performance to the ground-truth data. Moreover, Fig. 9 (middle, right) shows the corresponding results for different interest point detectors (Har-SIFT and MSER-SIFT) and on a different dataset (Oxford), which further validate this claim.

5.2.6 Experiment M06. Generalization

This experiment investigates whether the projections learnt on one dataset can improve matching performance on another dataset. We also test whether the projections trained on patches from one interest point detector can be used for another detector. Fig. 10 (Left) and (Middle)

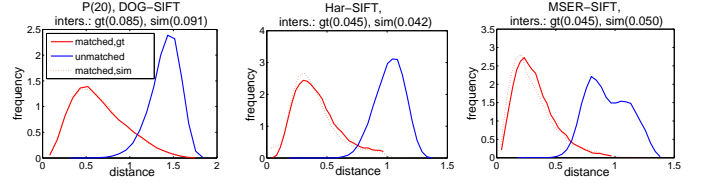


Fig. 9. *Experiment M05*. Comparison of the distance distribution of projected 20-dim features from simulated data (sim) with those from ground-truth data (gt). Left: P projected features on Liberty. Middle: Har-SIFT and Right: MSER-SIFT features on 15 image sequences from Oxford set. The intersection/union area ratio is given at the top of the figures for ground-truth data (gt) and simulated on (sim).

show the results for projections trained from one dataset (Notre-Dame, Yosemite) and tested on another one (Liberty). Note that there is no intersection between training and testing sets but the sets are likely to have similar distribution if sampled from the same scene. The results show that the performance is lower compared to the projections trained and tested on patches from the same scene (Liberty, Fig 8). However, the performance gain is larger for the Notre-Dame/Liberty test in Fig. 10 (Left) as these two datasets were obtained from structured building scenes, while Yosemite/Liberty score is lower as Yosemite was sampled from natural, well textured scenes. Finally, generalization from one interest point detector to another is also poor, in particular when compared with the PCA score. This is due to the difference in local structures detected by DoG and Harris, as the former responds to blobs and the latter to corners mainly.

Following the observation above, we conclude that the generalization property between different types of interest points and different types of scenes is weak. The projections generalize if the training and test set share similar statistics. While this may be seen as a disadvantage of the discriminant projections, the approach proposed in this paper provides a solution, as it allows

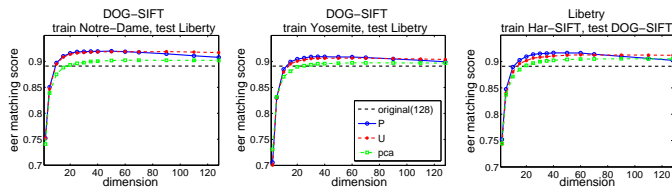


Fig. 10. *Experiment M06*. Generalization across different datasets and different interest point detectors. All the projections here are tested on Liberty with DoG-SIFT. The projections are trained for: Left: DoG-SIFT from Notre-Dame. Middle: DoG-SIFT from Yosemite. Right: Har-SIFT from Liberty.

to train the projections for any detector and any data without the ground truth.

5.2.7 Experiment M07. Oxford Dataset

In this experiment we further investigate the performance of discriminant projections on the Oxford matching benchmark from [31] using the area of precision-recall curves [15]. We test the performance using four different training strategies. The projections are first trained with ground-truth data (GT) on 15 training image sequences (50000 matched descriptor pairs) and tested on 3 test image sequences. Second strategy (TR) is to use the transformations estimated for each patch from the ground truth homography, to generate the matched patch with this transformation and to use these pairs for training the projections. The goal of the second strategy is to demonstrate how well local affine transformations can model the true transformation between the images if the true transformation parameters are known. Third approach (SIM) is to sample patches from the test images and to generate the simulated data with parameters estimated in experiment M03 for training the projections, which is possible in practical application scenario. Finally, fourth strategy (GT-Patch) is to use the projections trained on Liberty set provided by DoG detector and test on patches from Oxford data given by Harris detector. This experiment further tests the generalization across different data and detector.

As shown in Table 4, the projections result in higher scores than the original descriptor and the PCA. P gives slightly better performance than U , which is consistent with the observations in M05. Compared to GT, a small drop from 67.6% to 67.3% is observed for TR. This indicates that the simulated affine changes can model well the transformations in this dataset. Small decrease of performance results from noise and photometric variations which are not modeled by the affine transformation.

For the simulation approach, which can train the feature projections for any data and for any detector-descriptor combination, the performance is 66.4% (SIM). This is only 1.2% lower than for the ground-truth data (GT) but still 1.2% higher than for the original SIFT features and 6.9% higher than for the PCA projected features. It is important to note that this improvement

TABLE 4
Experiment M07. Average precision of the 40-dim projected features with various training strategies. See detailed explanation in the text.

	training strategies			
	GT	TR	SIM	GT-Patch
P (40)	67.6	67.3	66.4	54.6
U (40)	67.3	67.2	66.0	55.7
LDE (40)	64.9	65.4	66.1	–
SIFT (128): 65.2		PCA (40): 59.5		

has been achieved with dimensionality reduction from 128 to 40.

The result from projections trained and tested on different data and different detectors is 54.6% (GT-Patch), which is 11.8% lower than the result from SIM training. This demonstrates again the need for training data that has similar statistics to the test data and the usefulness of the simulation strategy.

In addition, we report results for LDE which considers only the nearest neighbors among matched and unmatched features. The intra-class covariance matrix estimated from all features rather than from nearest neighbors only (LDP) seems more stable and the projections result in higher performance.

5.2.8 Discussion.

In summary, the main observations from the experiments in this section are:

- Linear discriminative projections enable significant dimensionality reduction without compromising the performance and work better than PCA. P performs marginally but consistently better than U .
- The simulation is capable of generating the data very similar to the real one in terms of distance distributions, and performs nearly as well as the ground-truth data.
- The generalization properties of LDP are weak if the statistics of the data and feature detector differ significantly.
- Simulation parameters have an impact on the projection performance and have to be re-estimated if different features are used.

6 IMAGE RECOGNITION EXPERIMENTS

We further evaluate the LDP-projected feature in image recognition task. Image or object category recognition requires features that capture the dominant shape of the image structures and are insensitive to small appearance variations. We demonstrate on three standard recognition benchmarks that the proposed projection improves feature performance in this task. In the following we discuss the experimental setup and present the results for different datasets.

TABLE 5
Summary of recognition experiments.

Experiments	data		descriptor
	train	test	
<i>Experiment R01</i> . Table 6 simulation parameters	Scene-15 (100 images/cat.)	Scene-15	Dense-SIFT
<i>Experiment R02</i> . Table 7 recognition performance			
<i>Experiment R03</i> . Table 8 recognition performance	Caltech101 (15 images/cat.)	Caltech101	
<i>Experiment R04</i> . Table 9 recognition performance	VOC2007 (5011 images)	VOC2007	DoG-SIFT Dense-SIFT
<i>Experiment R05</i> . Table 10 generalization	Liberty	Scene-15	
	Scene-15 Caltech101 VOC2007	Scene-15	Dense-SIFT

6.1 Setup

6.1.1 Recognition System

We incorporate the proposed projections into a recognition system based on Spatial Pyramid Match Kernel (SPMK) [25]. Given a set of labeled training images the system extracts local regions and computes descriptors, from which it constructs a codebook with the k-means clustering ($k = 2000$). The image is first partitioned into L levels ($L = 1$) of increasingly fine location grids (1×1 , 2×2). Then a histogram of codeword occurrences is built for each location cell on each level. The similarity of two images is defined as a level-weighted sum of histogram intersections. A one-versus-all SVM classifier is then trained with these similarities. Given a query image the features are extracted, mapped to the codebook to build a multi-resolution histogram, which is then classified with the trained SVM.

6.1.2 Descriptors

Unlike in the matching experiment, we use uniform sampling of regions which was shown [32] to produce higher performance than the interest points. The features are extracted using regions of radius 16, 24, 32, 40, sampled uniformly every 8, 14, 20, 26 pixels, respectively. The regions are represented with SIFT descriptor, referred to as Dense-SIFT. 1.5×10^5 randomly selected regions ($\leq 5\%$) from all training regions are used to estimate projections with the simulating strategy. For each selected region, we simulate 9 regions with Gaussian distributed affine changes, which gives 10^6 matched features for the intra-class covariance estimation.

6.1.3 Data and Evaluation

Three commonly used image recognition benchmark datasets are used for evaluation. The experiments, datasets and the number of images used for training are summarized in Table 5.

Scene-15 dataset from [25], [24] contains fifteen scene categories each containing 200 to 400 images. Following [25] 100 images are used for training and the remaining ones for testing. The average recognition rates across categories are used for evaluation. All the experiments are repeated 10 times with different randomly selected training and testing images and the results averaged.

Caltech101 [29] consists of 101 object categories with 31 to 800 images per category. We run experiments 10 times with randomly selected 15 images for training, and

up to 50 remaining images for testing. The performance is measured in the same way as in Scene-15 experiment, with average recognition rates.

PASCAL VOC2007 [32] consists of 20 categories with 5011 training and 4952 test images. In contrast to Caltech101, each image in PASCAL database may contain multiple objects in various poses at different locations within the image, with background clutter and occlusion, which results in higher intra-class diversity than in Caltech101. The average precision (AP) as defined in [32] is used to measure the recognition performance.

6.2 Recognition Results

6.2.1 Experiment R01. Simulation Parameters

Since no point-to-point annotation exists in image/object categorization, the simulation strategy makes it possible to apply linear discriminant projections to local descriptors in this task. We first evaluate a number of combinations of different transformation parameters to find the best simulation parameters for this application. Table 6 shows standard deviations of Gaussian distributions from which the parameters were randomly sampled and the recognition results for different transform combinations on scene-15 dataset. The performance is little affected by different combinations tested in this experiment and it is always higher than PCA. SIM5 is the combination that gave the best results in matching experiments and it also optimizes the results in this test. Skewing and stretching changes have been discarded due to their limited effect on the performance. Similar tests were done for other datasets and the observations were consistent.

TABLE 6

Experiment R01. Transformation combinations and their performance on Scene-15. The number for each transform is the standard deviation of Gaussian distribution. SIM5 is the best combination from matching experiment *M03* with discarded skewing and stretching.

	SIM1	SIM2	SIM3	SIM4	SIM5	SIM6
rotate (degree)	3	5	8	12	7.5	11.2
scale (ratio)	1.1	1.3	1.4	1.7	1.13	1.18
translate (pixel)	1	2	5	10	4.8	6.7
$P(30)$, SPMK	83.4	83.9	84.5	84.2	84.6	84.0
SIFT (128): 83.5			PCA (30): 82.9			

6.2.2 Experiment R02. Scene-15

Table 7 displays the comparison of the average recognition rates for the original SIFT, PCA- and P -projected features. P -projected features outperform the original SIFT by 1.1% and PCA-projected by 1.7%. Our implementation of SPMK with 128-dim SIFT gives 83.5% which is higher than the performance of this approach reported in the original paper [25] (81.4%) mostly due to larger size codebook. The best performance of 83.7% on this data was achieved in [38] where a probabilistic Latent Semantic Analysis is applied to SPMK model (SP-pLSA). Our P -projected features with 30 dimensions give 84.6% and exceed the performance of all other state-of-the-art approaches. The recognition system may benefit from the projections and reduced dimensionality at the codebook construction stage or feature to codebook assignment. It is however not straightforward to identify the most beneficial part of the system as the intermediate results cannot be directly interpreted.

TABLE 7

Experiment R02. Comparison of the average correct rate on 10 runs with 100 training images per category on Scene-15.

P (30)	SIFT (128)	PCA (30)	[25]	[37]	[21]	[38]
84.6	83.5	82.9	81.4	83.3	77.0	83.7

6.2.3 Experiment R03. Caltech101

Table 8 shows the mean recognition rates for Caltech101 data averaged for 10 runs. Our method is a single kernel approach with gray-value SIFT only on dense sampled regions, we therefore only compare our results with the state-of-the-art single-kernel approaches in Table 8. The highest score of 63.2% was reported in [40], in which an image specific distance metric is learned with paired constraints. The P -projected 30-dim features outperform this approach by 1.7%. The gain with respect to the original SIFT and PCA-projected features is small. With many variants of the recognition methods based on local descriptors and SIFT in particular, the results reported in the literature are saturated and further improvements require more diverse measurements from the image.

TABLE 8

Experiment R03. Comparison of the average correct rate on 10 runs with 15 training images per category on Caltech101.

P (30)	SIFT (128)	PCA (30)	[44]	[43]	[25]	[45]
64.9	64.5	64.0	52.0	50.0	56.4	59.1
			[38]	[40]	[30]	[39]
			59.8	63.2	61.0	60.5

6.2.4 Experiment R04. PASCAL VOC2007

The Average Precisions (AP) from 20 object categories from this benchmark are displayed in Table 9. The results show that the P -projected 30-dim features yield a gain

TABLE 9

Experiment R04. Comparison of the average precision on PASCAL VOC2007.

cat.	P (30)	SIFT	PCA (30)	[41]	C-SIFT [46]	[42]
cat.	72.2	70.3	71.7	65.0	59.9	64.3
plane	55.7	55.6	53.6	44.3	43.5	52.0
bicycle	46.3	40.6	38.9	48.6	37.1	45.0
bird	65.9	63.7	63.1	58.4	55.5	60.4
boat	26.8	26.1	20.0	17.8	20.7	20.3
bottle	59.0	55.3	56.7	46.4	34.4	49.2
bus	73.7	71.9	71.1	63.2	54.9	69.3
car	55.3	54.7	53.4	46.8	36.7	47.9
cat	51.1	49.2	48.7	42.2	46.2	49.4
chair	36.2	36.6	35.9	29.6	27.8	34.0
cow	46.1	46.6	43.9	20.8	39.2	37.5
table	38.9	40.4	38.0	37.7	29.8	40.0
dog	72.2	72.7	72.7	66.6	66.9	71.2
horse	60.9	58.0	57.2	50.3	43.1	57.7
bike	80.9	80.1	80.0	78.1	78.5	80.6
person	25.7	24.0	23.8	27.2	31.0	32.6
plant	36.4	35.1	33.3	32.1	41.5	35.7
sheep	47.7	43.0	42.7	26.8	32.4	42.6
sofa	75.5	71.3	70.8	62.8	61.7	68.0
train	47.0	45.6	43.9	33.3	35.5	47.7
monitor						
MAP	53.7	52.0	51.0	44.9	43.9	50.2

of 2.7% in terms of Mean AP (MAP) over the PCA-projected ones and 1.7% over original 128-dim SIFT. It is also observed that P outperforms PCA on 19 categories and original SIFT on 16 categories.

We also compare our scores with the state-of-the-art results. As demonstrated in Table 9, the P -projected features obtain the best performance for 12 out of 20 categories and exceed the best reported result [42] by 3.5%. The improvement is even higher compared to the other two approaches [41], [46]. It is worth noting that all these methods combine color-based descriptors with SIFT resulting in high dimensional features, e.g., 384 dimensions for the best color sift (C-SIFT) in [46]. Therefore, the significant advantage of our approach is not only the performance improvements but also much less memory and computational time requirements, which is crucial for large scale computer vision systems.

In addition, we also test on PASCAL VOC2008 Challenge [32] by training on 2111 training images and testing on 2221 validation images, since ground truth for test set is not available. Consistent improvement is observed. P -projected 30-dim features, PCA-projected 30-dim features and 128-dim SIFT yield 41.3%, 40.8% and 39.7% in terms of MAP. The P -projected 30-dim features also outperform the best single kernel score of 38.8% [36] with 384-dim color-SIFT[46], which produced the best result on this dataset. This confirms the superiority of our P projection in both performance and scalability.

6.2.5 Experiment R05. Generalization

In this experiment the performance of the projections is compared by training and testing on different datasets. As shown in Table 10, the lowest performance is given by the projections trained on patch dataset from matching experiments in section 5. This is due to: 1) patches in this dataset are sampled from one scene type only

TABLE 10

Experiment R05. SPMK recognition with descriptor projections trained and tested with different datasets. All the experiments are tested on Scene-15 with the projections trained on different datasets.

	training set			
	Scene-15	Patch	Caltech101	VOC2007
P (30)	84.6	82.9	83.9	83.8
SIFT (128): 83.5		PCA (30): 82.9		

(Liberty); 2) uniform sampling that is used in recognition task results in more diverse local structures than from the interest point detector used for the patch database. Projections trained from Caltech101 or VOC2007 give better scores as the sampling strategy and the data is more similar to the test set. However, the highest performance is still obtained by training and testing on the sets from the same distribution. This further validates the observations from section 5.2 that the projections should be trained on the data as close to the test distribution as possible.

7 DISCUSSION AND CONCLUSIONS

In this paper, we have presented and evaluated LDP approach and put it in the context of other discriminant projections. A number of dimensionality reduction techniques have been discussed with analytical and experimental in depth analysis of the LDP.

We have applied LDP to state-of-the art descriptor SIFT as well as normalized grayvalue patches, but the approach can be used for any type of descriptors. We have demonstrated that LDP can reduce the dimensionality without compromising the performance and outperforms the commonly used PCA. This makes it a useful tool for computationally demanding large-scale image recognition and retrieval problems.

We have shown that LDP-projected 30-dim features combined with SPMK based recognition system improve upon the original SIFT, PCA-projected features and also state-of-the-art results for three recognition benchmarks. This suggests that modeling of geometric changes gives useful projections in the categorization problem, even though the main changes are in the appearance/shape.

We have proposed a simulation based training method and demonstrated its usefulness in two typical image descriptor applications. The results show that the simulated data distribution is similar to the real one and can be used instead. It is a significant contribution as it makes it possible to apply the LDP as well as other discriminant projection approaches to the problems where point-to-point correspondence ground truth is not available.

We have not been able to find a general projection that would improve SIFT for any task despite many experiments with different variants of projections, descriptors and training methods. It seems that SIFT can be made more compact, but better results have been achieved

only by making the projections specific to the data. Interestingly, this also suggests that the standard benchmarks used in this paper have some specific properties and very high performance on one may not generalize to the other. As with any discriminative approach there is a risk of overfitting the projections but our simulating approach can provide sufficient number of data points for training and it allows to control their variability. The improvements are small but consistent in all experiments. The data, projection matrices and software libraries are made available at [31] and may serve as a benchmark for future projections and descriptors.

In the future, we will investigate discriminant projections for other descriptor based applications i.e., image retrieval from large databases. Another possible direction is to incorporate means of feature clusters into the cost function being optimized. Non-linear projection methods based on kernels, including kernel fusion of various descriptor types are also of interest.

APPENDIX

In this section we demonstrate that projections P and U defined in (6) and (5) have equivalent orientations.

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

$$C_S^{-\frac{1}{2}}C_D C_S^{-\frac{1}{2}} \cdot R = R \cdot \Omega \quad (12)$$

If equation (12) is multiplied by $C_S^{-\frac{1}{2}}$, then we obtain:

$$C_S^{-1}C_D \cdot C_S^{-\frac{1}{2}}R = C_S^{-\frac{1}{2}}R \cdot \Omega \quad (13)$$

where $C_S^{-\frac{1}{2}}R$ are the eigenvectors of $C_S^{-1}C_D$, which means that $U = \text{eig}(C_S^{-1}C_D) = C_S^{-\frac{1}{2}}R = P$ if only orientations are considered. This equivalence is also verified by the theorem of simultaneous diagonalisation of the covariance matrices. Let Y^P be the projected feature spaces: $Y^P = \{P^T x | x \in X\}$.

$$\begin{aligned} C_S|_{Y^P} &= \text{cov}(P^T \cdot (x_i - x_j))_{(i,j) \in S} = P^T \cdot C_S \cdot P \\ &= R^T C_S^{-\frac{1}{2}} C_D C_S^{-\frac{1}{2}} R \end{aligned} \quad (14)$$

Since R is an orthogonal matrix, then

$$C_S|_{Y^P} = I \quad (15)$$

This property is illustrated by the projected 2-dim data points in Fig. 1 where each small cluster is circular.

$$\begin{aligned} C_D|_{Y^P} &= \text{cov}(P^T \cdot (x_i - x_j))_{(i,j) \in D} = P^T \cdot C_D \cdot P \\ &= R^T C_S^{-\frac{1}{2}} C_D C_S^{-\frac{1}{2}} R \end{aligned} \quad (16)$$

R is the matrix of all eigenvectors of $C_S^{-\frac{1}{2}}C_D C_S^{-\frac{1}{2}}$, thus

$$C_D|_{Y^P} = \Omega \quad (17)$$

From the above equations, P is a linear transformation that diagonalizes two symmetric matrices C_S and C_D simultaneously. According to the theorem in [27] (p.31),

Ω and P are the eigenvalue and eigenvector matrices of $C_S^{-1}C_D$, which demonstrates that P has the same orientations as U .

ACKNOWLEDGMENTS

This research was sponsored by the BBC and the EPSRC UK project EP/F003420/1. J. Matas was supported by Czech Ministry of Education Research Project MSM6840770038.

REFERENCES

- [1] A. C. Berg and J. Malik, "Geometric blur for template matching," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 1-607-1-614, 2001.
- [2] S. Yan, D. Xu, B. Zhang, H.-J. J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40-51, 2007.
- [3] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319-2323, 2000.
- [4] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323-2326, 2000.
- [5] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems*, vol. 14, pp. 585-591, 2001.
- [6] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems*, 2003.
- [7] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information," in *Advances in Neural Information Processing Systems*, vol. 15, pp. 505-512, 2002.
- [8] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 846-853, 2005.
- [9] K. Q. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in Neural Information Processing Systems*, vol. 18, pp. 1473-1480, 2006.
- [10] S. Winder, G. Hua and M. Brown, "Picking the best DAISY," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 178-185, 2009.
- [11] G. Hua, M. Brown, and S. Winder, "Discriminant embedding for local image descriptors," in *Proc. IEEE Int'l Conf. Computer Vision*, pp. 1-8, 2007.
- [12] K. Mikolajczyk and J. Matas, "Improving descriptors for fast tree matching by optimal linear projection," in *Proc. IEEE Int'l Conf. Computer Vision*, pp. 1-8, 2007.
- [13] H. Cai, K. Mikolajczyk and J. Matas, "Learning Linear Discriminant Projections for Dimensionality Reduction of Image Descriptors" in *Proc. British Machine Vision Conf.*, 2008.
- [14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int'l J. Computer Vision*, vol. 60, no. 2, pp. 91-110, November 2004.
- [15] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615-1630, 2005.
- [16] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. Van Gool, "A comparison of affine region detectors," *Int'l J. Computer Vision*, vol. 65, no. 1/2, pp. 43-72, 2005.
- [17] E. Tola, V. Lepetit, and P. Fua, "A fast local descriptor for dense matching," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1-8, 2008.
- [18] V. Lepetit and P. Fua, "Keypoint recognition using randomized trees," in *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1465-1479, 2006.
- [19] Y. Ke and R. Sukthankar, "Pca-sift: a more distinctive representation for local image descriptors," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 506-513, 2004.
- [20] K. Grauman and T. Darrell, "The pyramid match kernel: discriminative classification with sets of image features," in *Proc. IEEE Int'l Conf. Computer Vision*, vol. 2, pp. 1458-1465, 2005.
- [21] J. C. van Gemert, J. M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, "Kernel codebooks for scene categorization," in *Proc. European Conf. Computer Vision*, 2008.
- [22] T. Tuytelaars and K. Mikolajczyk, "Local Invariant Feature Detectors: A Survey", in *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177-280, 2008.
- [23] W. Triggs, "Detecting keypoints with stable position, orientation, and scale under illumination changes.", in *European Conf. Computer Vision*, pp. 100-113, 2004.
- [24] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 524-531, 2005.
- [25] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2169-2178, 2006.
- [26] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning distance functions using equivalence relations," in *Proc. Int'l Conf. Machine Learning*, pp. 11-18, 2003.
- [27] K. Fukunaga, *Introduction to statistical pattern recognition (2nd ed.)*. Academic Press Professional, Inc., 1990.
- [28] J. Matas, O. Chum, U. Martin, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. British Machine Vision Conf.*, vol. 1, pp. 384-393, 2002.
- [29] L. F. Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *Computer Vision and Pattern Recognition Workshop on Generative Model Based Vision*, pp. 178-178, 2004.
- [30] P. Jain, B. Kulis, and K. Grauman, "Fast image search for learned metrics," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1-8, 2008.
- [31] Data, feature detectors, evaluation protocols. <http://www.featurespace.org>.
- [32] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge Results," <http://www.pascal-network.org/challenges/VOC/>.
- [33] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighborhood component analysis," in *Advances in Neural Information Processing Systems*, 2004.
- [34] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," in *Advances in Neural Information Processing Systems*, 2004.
- [35] L. Yang and R. Jin, "Contents distance metric learning: A comprehensive survey," Technical report, Department of Computer Science and Engineering, Michigan State University, 2006.
- [36] M. A. Tahir, K. van de Sande, J. Uijlings, F. Yan, X. Li, K. Mikolajczyk, J. Kittler, T. Gevers, and A. Smeulders, "Surreyvuva_srkd method," in *Pascal VOC 2008 Workshop*, 2008.
- [37] J. Liu and M. Shah, "Scene modeling using co-clustering," in *Proc. IEEE Int'l Conf. Computer Vision*, pp. 1-7, 2007.
- [38] A. Bosch, A. Zisserman and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," in *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712-727, 2008.
- [39] S. Fidler, M. Boben and A. Leonardis, "Similarity-based cross-layered hierarchical representation for object categorization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1-8, 2008.
- [40] A. Frome, Y. Singer, F. Sha and J. Malik, "Learning Globally-Consistent Local Distance Functions for Shape-Based Image Retrieval and Classification," in *Proc. IEEE Int'l Conf. Computer Vision*, pp. 1-8, 2007.
- [41] G. Wang, D. Hoiem, and D. Forsyth, "Learning image similarity from flickr groups using stochastic intersection kernel machines," in *Proc. IEEE Int'l Conf. Computer Vision*, 2009.
- [42] F. Shahbaz Khan, J. van de Weijer, and M. Vanrell, "Top-down color attention for object recognition," in *Proc. IEEE Int'l Conf. Computer Vision*, 2009.
- [43] K. Grauman and T. Darrell, "The pyramid match kernel: efficient learning with sets of features," *J. Machine Learning Research*, vol. 8, no. 4, pp. 725-760, 2007.

- [44] A. C. Berg, T. L. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondences," in *Pro. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 26–33, 2005.
- [45] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "Svm-knn: Discriminative nearest neighbor classification for visual category recognition," in *Pro. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2126–2136, 2006.
- [46] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, in press, 2010.

PLACE
PHOTO
HERE

Hongping Cai received the BS degree and MS degree in Computational Mathematics in 2001 and in 2003 at National University of Defense Technology, China. She is currently a PhD student in College of Electronic Science and Engineering at National University of Defense Technology, China. During her PhD study, she spent one year and nine months as a visiting student in Centre for Vision, Speech and Signal Processing, University of Surrey. Her research interests include local image feature, dimensionality reduction, vision recognition and retrieval.

PLACE
PHOTO
HERE

Krystian Mikolajczyk is a Senior Lecturer in Robot Vision at the Centre for Vision, Speech and Signal processing at the University of Surrey, UK. He did his PhD at INRIA Grenoble (France) on invariant interest points and then held post-doc positions at INRIA, University of Oxford (UK) and Technical University of Darmstadt (Germany), working primarily on image recognition problems. His main scientific contributions are in the domain of invariant image descriptors for matching and recognition.

He currently leads a group of PhD students and post-docs focused on visual recognition problems including issues like image matching, categorization, retrieval as well as object and action recognition. <http://personal.ee.surrey.ac.uk/Personal/K.Mikolajczyk> Centre for Vision, Speech and Signal Processing, University of Surrey, UK.

PLACE
PHOTO
HERE

Jiri Matas received the MSc degree in cybernetics (with honors) from the Czech Technical University, Prague, Czech Republic, in 1987 and the PhD degree from the University of Surrey, UK, in 1995. Since 1997, he has been with the Center for Machine Perception at the Czech Technical University. He has published more than 150 papers in refereed journals and conferences.

His publications have more than 3000 citations in the ISI Thomson-Reuters Science Citation Index, his SCI H-index is 21. He received the best paper prize at the British Machine Vision Conferences in 2002 and 2005 and at the Asian Conference on Computer Vision in 2007. J. Matas has served in various roles at major international conferences (e.g. ICCV, CVPR, ICPR, NIPS, ECCV), co-chairing ECCV 2004 and CVPR 2007. He is on the editorial board of IJCV and IEEE T. PAMI.

His research interests include object recognition, image retrieval, sequential pattern recognition, ensemble methods, invariant feature detection, and Hough Transform and RANSAC-type optimization.