# PER-PATCH METRIC LEARNING FOR ROBUST IMAGE MATCHING

*Sezer Karaoglu, Ivo Everts, Jan C. van Gemert, and Theo Gevers*

Intelligent Systems Lab, Amsterdam, University of Amsterdam, 1098 XH Amsterdam, The Netherlands

## ABSTRACT

We propose a patch-specific metric learning method to improve matching performance of local descriptors. Existing methodologies typically focus on invariance, by completely considering, or completely disregarding all variations. We propose a metric learning method that is robust to only a range of variations. The ability to choose the level of robustness allows us to fine-tune the trade-off between invariance and discriminative power.

We learn a distance metric for each patch independently by sampling from a set of relevant image transformations. These transformations give a-priori knowledge about the behavior of the query patch under the applied transformation in feature space. We learn the robust metric by either fully generating only the relevant range of transformations, or by a novel direct metric. The matching between query patch and data is performed with this new metric.

Results on the ALOI dataset show that the proposed method improves performance of SIFT by $6.22\%$ for geometric and $4.43\%$ for photometric transformations.

## 1. INTRODUCTION

Viewing and lighting condition changes in real-world scenes cause substantial variations in image feature representations. Significant progress has been made in developing image representations that are invariant to transformations such as photometry [1, 2] or geometry [3, 4, 5, 6]. Image representations invariant to such changes are beneficial for applications such as object recognition, image retrieval and scene recognition.

A full invariant representation, unfortunately, leads to a decrease in discriminative power [7]. This drawback is due to distinguishing transformations that a full invariant representation cannot capture. For example, under rotational invariance a "6" is identical to a "9", and under shading invariance the texture of "grass" turns into "moss". Another disadvantage of invariant image representations is that they negatively influence stability [8, 9, 10]. This is due to their sensitivity to noise when the image signal is low or ambiguous. For example, a rotational invariant based on the dominant orientation [5] becomes unstable when multiple equally dominant orientations are present. Illumination invariant representations based on intensity normalization such as normalized-$rgb$ or $hue$ [2] become unstable for low intensity values.

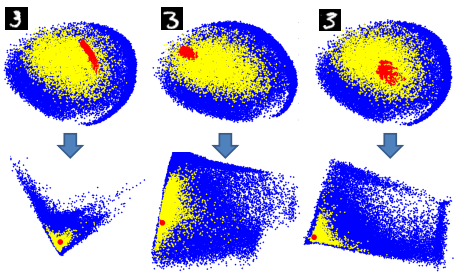

**Fig. 1**. 2D PCA projection of SIFT extracted from dataset samples (blue); 1000 affine transformations of the top-left image patch (red); same-class samples (yellow). The top row is the original space, the bottom row is after learning the metric.

Current invariant methods are always "on". One can either choose to use the invariance, or choose not to use it. There is no middle-ground. It is not possible to have invariance for only some shading or only slight rotations. These rigid properties of current invariants play a central role in the trade-off between invariance and discriminative power. Here, we propose to replace these binary on/off invariants, by steering the invariance to a limited range of disturbances. Such a limited degree of invariance is called *robustness*. For example, in the case of rotation, the proposed method can determine that a "6" is only invariant up to $\pm 45^o$ of (and thus robust to) rotation, therefore eliminating the confusion with "9".

By allowing a degree of invariance, a single global image representation cannot be used as it depends on the specific image content how the limited transformation range will take effect, which is illustrated by the "6" and "9" example. Therefore, the proposed method is required to achieve robustness on a per-patch basis. Fig. 1 (top) illustrates that feature distributions after a transformation depend on the patch content, since even instances within the same class behave differently (red versus yellow). The bottom row of Fig. 1 illustrates the effect of steerable invariance applied to each patch.

To achieve robustness, we compute a Mahalanobis metric for each individual patch. In effect, the metric weights the subset of feature dimensions that require robustness. For this, a relevant subset of transformations is generated and a metric that is specific for only those transformations is learned. We present two approaches for learning the metric: (i) *full*

and (ii) *direct*. The *full* method generates synthetic image patches, extracts descriptors for each patch, and obtains robustness through a metric that is learned on these descriptors. In the *direct* approach, we generate a transformation map only once, and use this map to directly estimate the metric from the patch without explicitly generating any synthetic images.

## 2. RELATED WORK

Approaches that aim to achieve (full) invariance either use a transformation model based on the laws of physics [1, 2] or a model of the observed variations [3, 4, 11, 5]. The two disadvantages of invariants, stability and discriminative power, can be addressed by propagating camera noise parameters [8] or by deriving quasi-invariants [9]. Noise propagation requires proper noise estimation and the quasi-invariants are incomparable over different images and thus cannot be used for matching.

In contrast to employing pre-determined models, (deep) learning methods learn invariant features from unsupervised training examples [12, 13, 14]. Such methods do not explicitly model invariance as they attain robustness from training examples. Therefore, learning methods require large amounts of training data which is hard to obtain. Moreover, learning approaches do not directly incorporate known physical laws of the world. In this work, we use a hybrid approach of modeling robustness by learning from synthetically generated geometric and photometric data.

Synthetically generated data can be used to directly create variation in the train and test samples [15, 16, 17]. Other brute-force methods like ASIFT [18] generate a full range of affine transformations for both training and testing images which are used in an exhaustive matching scheme. Similar to our work, Simard et al. [19] avoid brute-force approaches and use synthetically generated images to learn a robust distance metric which is tangent to the manifold that is spanned by the generated transformations. We also learn a robust metric, however, where Simard et al. [19] require pixel values to estimate a manifold, our method estimates a Mahalanobis distance, which is applicable to any feature representation such as SIFT. To improve discriminability of a local descriptor, Cai et al [15] also propose to learn a projection matrix for a limited range of affine transformations through generated data. However, it is important to note that the authors learn a global projection whereas this paper proposes to learn patch specific projections. Fig. 1 illustrates that the same transformations applied to even the same class instances has different effects for different patches. These variations are thus patch-dependent and might not reflect the appropriate effect on other patches from the same class.

This paper has following contributions: (i) we demonstrate that a full invariant representation leads to a decrease in the discriminative power of a descriptor. Accordingly, we propose to limit the degree of invariance and augment the discriminativeness of a descriptor. (ii) we demonstrate that a single global image representation cannot be used as the effect of transformation essentially depends on the specific image content. Thus, a patch specific metric is proposed. (iii) we propose two alternatives to learn per-patch metric: by either explicitly applying transformations or obtaining directly from the patch.

## 3. PER-PATCH METRIC LEARNING

We first develop a metric that is learned from synthetically generated transformations of photometric and geometric distortions.

**Geometric Transformations.** Images are subject to geometric distortions introduced by perspective effects caused by view point changes. For small patches, the perspective transformation $(x', y')^T$ can be approximated by an affine transformation for a given point $(x, y)^T$ as

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} s_x \cos\alpha & -\tau_x \sin\alpha & t_x \\ \tau_y \sin\alpha & s_y \cos\alpha & t_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}, \quad (1)$$

where $s$ denotes scale, $\tau$ represents shearing, $\alpha$ is the rotation angle, and $t$ denotes translation. Sample generation involves repetitive random selection from appropriate parameter ranges.

**Photometric Transformations.** To model photometric changes, we assume Lambertian reflection. Accordingly, the color response ($I$) for the visible spectrum ($\lambda$), using a camera with spectral sensitivity ($f$) and an illumination source with the spectral power distribution ($e$) can be defined as

$$I = \vec{n}.\vec{s} \int_{\lambda} e(\lambda)\rho(\lambda)f(\lambda)d\lambda. \quad (2)$$

$s$, $n$ and $\rho$ denote the illumination direction, the surface normal and the surface albedo respectively. To obtain photometric robustness, we generate variations caused by Lambert's Law. For the same surface patch the viewpoint and illuminant spectral power distribution are the same. The color response can only vary due to changes in illumination direction (i.e. $\rho$, $f$, $e$ and $n$ remain constant). Therefore, the changes in $I$ can be modeled by illuminating the patch from different positions (See Fig. 2).

The center of the image patches are considered to be at $(0, 0, 0)$ and placed perpendicular to the light source position. Then, the light source position is systematically sampled within a certain radius. The patches are sufficiently small to be assumed planar. Hence, surface normals are equal for the patch under consideration. Thus, the light source direction is the only factor determining the effect of the photometrical changes.

### 3.1. Metric Learning

A Mahalanobis distance metric between image features $\mathbf{x}_i$ and $\mathbf{x}_j$ is parameterized by the matrix $M$,

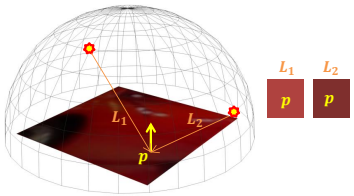


**Fig. 2**. Illustration of photometric changes. The color response of point $p$ changes due the angle between the incident light ($L$) and surface normal (yellow arrow) at the point $p$. According to Lambertian law, if this angle becomes smaller, the color response becomes brighter.

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T M (\mathbf{x}_i - \mathbf{x}_j)}, \tag{3}$$

where $M$ is a positive semi-definite matrix. A straightforward approach to compute $M$ is to use $M = C^{-1}$ where $C$ is the empirical feature covariance of the training data [20, 21]. The rationale behind using the inverse covariance as a metric is that a high variance in a feature dimension means that this dimension is not very stable. The most informative dimensions are those that have low variances.

For large-scale matching it is convenient to use fast indexing techniques such as trees [22]. Such techniques typically work with the Euclidean distance. To this end, the metric can be rewritten to

$$(\mathbf{x}_i - \mathbf{x}_j)^T W^T W (\mathbf{x}_i - \mathbf{x}_j) = ||W\mathbf{x}_i - W\mathbf{x}_j||^2, \tag{4}$$

where $W = M^{-\frac{1}{2}}$. This effectively scales the feature space with an affine transformation $W$ to allow the Euclidean distance to be used for metric $M$.

Robustness is obtained by generating a limited range of transformations for a single patch. From these generated samples $C$ is estimated, which is then used to compute $M$.

We coin this simple method for metric learning the *full* approach, since it needs to fully generate a large number of transformations for a patch in order to extract the image features and estimate the covariance matrix.

### 3.2. Direct Metric Estimation

Instead of using the brute-force approach of computing $C$ by explicitly generating patches and extracting features from them, we propose a *direct* approach to estimate the metric per-patch. In the direct approach, the covariance is estimated from a single patch.

Let ($\mathbf{x}$ be an image feature vector as a column, with $N$ dimensions. For clarity, we start with pixel values to explain the direct metric estimation, i.e. $\mathbf{x}$ is a vector of pixels. Later we show our method readily applies to other descriptors.

**Pixel Values.** The direct metric is a combination of two terms, a transformation probability $D_T$ that a pixel moves to a different position after transformation $T$, and the patch-specific covariance term $V$ of the feature values.

Let $D_T$ be a symmetric matrix of size $N$x$N$, containing the probabilities $P_T(i|j)$ of a pixel at position $j$ affecting the position of pixel $i$ under a transformation $T$. Note that this transformation is independent of the actual feature values. Matrix $D_T$ represents the per-pixel transformation probability which is determined by simulating a large set of transformations and comparing the transformed patch with the ground truth location obtained through the homography in eq 1.

The matrix $V$ of size $N$x$N$ is the variance matrix with elements $\sigma(i,j)$ representing the covariance of pixel values at position $i$ with respect to position $j$. To compute the covariance matrix, we need the expected average weighted pixel value $\mathbf{x_i}$ at position $i$ after the transformation $T$, which is given by:

$$\mathrm{E}[\mathbf{x}_T(i)] = \sum_{j=0}^{N^2} P_T(i|j)\mathbf{x}(j) = D_T\mathbf{x}^T, \tag{5}$$

where $\mathbf{x}_T$ represents the pixel vector $\mathbf{x}$ after the transformation $T$. The expected value of the transformed image $\mathbf{x}_T$ is denoted by $\mathrm{E}[\mathbf{x}_T(i)]$ and represents the average image of all transformations that are present in $D_T$. The covariance $\sigma(i,j)$ after transformation $T$ is then

$$\sigma(i,j) = \mathrm{E}\big[\big(\mathrm{E}[\mathbf{x}_T(i)] - \mathbf{x}_T(i)\big)\big(\mathrm{E}[\mathbf{x}_T(j)] - \mathbf{x}_T(j)\big)\big]. \tag{6}$$

Note that in the transformation of $\mathbf{x}_T(i)$ and $\mathbf{x}_T(j)$ it is allowed for pixels to move independently to other pixels. Eq 6 can be rewritten in matrix form to obtain $V$ directly

$$W_V = [[D_T > 0]] \bullet \big(D_T\mathbf{x}\mathbf{1}^T - (\mathbf{x}^T\mathbf{1})^T\big), \tag{7}$$

$$V = \frac{1}{N^2} D_T \bullet W_V W_V^T, \tag{8}$$

where $\mathbf{1}$ is a column vector of all ones, $\bullet$ denotes element-wise multiplication and $[[\cdot]]$ indicate Iverson brackets which resolves a (matrix) element to 1 when the argument is true, and 0 otherwise. The metric is computed by $M = V^{-1}$, and the transformation by $W = V^{-\frac{1}{2}}$.

**SIFT.** For other descriptors, the $D_T$ matrix of transformation probabilities can be reused and is not required to be recomputed. The $V$ matrix, however, has to be adapted to the specific form of the descriptor. In the case of SIFT, $D_T$ is converted to $D_T^{\text{sift}}$ of size 128x128. In contrast to generating all possible sift values, $D_T^{\text{sift}}$ has to be computed once only.

The 128 dimensions of SIFT comprise of a 4x4 spatial grid and 8 angular bins (4x4x8). We use the pixel-based transformation probabilities $D_T$ to directly calculate the probability $P_T^{\text{sift}}(i|j)$ for spatial SIFT bins $i$ and $j$ with

$$P_T^{\text{SIFT}}(i|j) = \sum_{y=0}^{N^2} [[f(x) = i]][[f(y) = j]] P_T(x|y), \tag{9}$$

where the function $f(x)$ maps the pixel at location x to the correct SIFT-bin $i$. For the angular transformation probabilities, we assume independence with the spatial bins. The unit

circle is sampled with 360 vectors and transformations are applied to these vectors. Since the original orientation is known, this results in counting how often a vector switches bins after the transformations. The joint 128x128 matrix $D_T^{\text{sift}}$ is calculated by multiplying the angular probabilities with the spatial probabilities.

## 4. EXPERIMENTS

**Dataset and Implementation Details.** The *full* and *direct* metric methods are evaluated on the ALOI dataset [23] for SIFT descriptor robustness against geometric and photometric distortions. The ALOI dataset contains 1000 objects under varying imaging conditions. We use variations due to camera viewpoint and illumination direction as geometric and photometric distortions respectively. To annotate matching pairs, the same procedure is followed as in [24]. In total 8300 matching pairs are extracted of which 200 are used for validation. Classification is performed by feature matching in a 1-NN classification scheme.

### 4.1. Geometric Robustness

First, we evaluate per-patch metric learning under geometric distortions. The performance of the proposed methods is compared against the original SIFT and other methods.

The optimal parameter ranges are obtained on the validation set. Parameters are repetitively sampled (1500 times) from various ranges to either generate and apply geometric transformations in the *full* approach or to estimate the transformation probabilities $P_T(i|j)$ in the *direct* approach. A joint optimization of parameters on the validation set yields a translation in [-2:2] and shearing in [-.1:.1] for the SIFT descriptors. Scale and orientation do not affect the performance as the positional difference between viewpoints does not yield large scale and orientation variations.

The results in Table 1 show that the proposed *full* and *direct* methods have a significant improvement of $6.57\%$ and $6.22\%$ over the SIFT performance respectively. The significance is validated by t-test a ($p < 0.001$). The substantial performance increase is due to the fact that viewpoint variations degrade the SIFT performance for matching. Considering full-rotation invariance for SIFT leads to a dramatic performance loss as the most discriminative information is ignored and due to the visual complexity it is harder to estimate a dominant gradient orientation.

Additionally, we evaluate raw pixels and tangent distance ($TD$) [19] on this set. We obtain a classification rate of 16.54% and 23.47% respectively. As discussed in Section 2, Simard et al. invoke prior information by generating small global transformations. The performance improvement over raw pixel values supports our idea of exploiting prior information for steering the invariance. However, as expected, the raw pixel matching performance is far beyond the SIFT performance which makes $TD$ less applicable when it is necessary to use features except raw pixel values.

| SIFT | +Rot. Inv. | Proposed *full* | Proposed *direct* |
|---|---|---|---|
| 54.85% | 13.86% | **61.42%** | 61.07% |

**Table 1**. Matching performance on geometric distortions. Left to right: original SIFT descriptor, full-rotation invariant SIFT descriptor, proposed *full*(synthetic) and *direct* methods. Note that full-rotation invariance is unstable whereas limited range of invariance is stable.

### 4.2. Photometric Robustness

We evaluate the proposed *full* and *direct* metric methods for photometric robustness. The *full* method operates by generating synthetic samples as explained in *Photometric transformations*. The results are shown in Table 2. The proposed *full* method significantly outperforms the SIFT performance (t-test with $p < 0.001$). Moreover, the *direct* method does not rely on synthetic photometric data generation and it outperforms the SIFT as well. Thus, the direct approach, which is developed for achieving robustness against geometric transformations, is also applicable in the context of photometric variations.

| SIFT | Proposed *full* | Proposed *direct* |
|---|---|---|
| 73.4% | **77.86%** | 76.88% |

**Table 2**. Matching performance on photometric distortions. Note that both methods perform similar, where "direct" is much faster.

## 5. CONCLUSION

In this paper, we propose a generic patch-specific robust metric learning method to improve matching performance of local descriptors. We show that a full invariant representation leads to a decrease in discriminative power of descriptors. Therefore, we propose a per-patch metric learning method that is invariant to only a range of variations. We propose to learn a patch specific a Mahalanobis metric. Two approaches for learning the metric are presented: (i) *full* and (ii) *direct*. The proposed approaches are validated on ALOI dataset for two different image transformations, namely, geometric and photometric. It has been shown that the proposed approaches outperform the original SIFT descriptor matching performance.


# Acknowledgment

This research was supported by the Dutch national program COMMIT.

## 6. REFERENCES


[1] J. Geusebroek, R. van den Boomgaard, A. Smeulders, and H. Geerts, "Color invariance," *TPAMI*, 2001.

[2] T. Gevers and W. M. A. Smeulders, "Color based object recognition," *Pattern recognition*, 1999.

[3] A. Baumberg, "Reliable feature matching across widely separated views," in *CVPR*, 2000.

[4] T. Lindeberg, "Feature detection with automatic scale selection," *IJCV*, 1998.

[5] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *IJCV*, 2004.

[6] L. Zhang, Z. Lei, Z. Guo, and D. Zhang, "Monogenic-lbp: A new approach for rotation invariant texture classification.," in *ICIP*, 2010.

[7] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *ICCV*, 2007.

[8] T. Gevers and H. Stokman, "Robust histogram construction from color invariants for object recognition," *TPAMI*, 2004.

[9] J. van de Weijer, T. Gevers, and J. M. Geusebroek, "Edge and corner detection by photometric quasi-invariants," *TPAMI*, 2005.

[10] I. Everts, J. C. van Gemert, T. E. J. Mensink, and T. Gevers, "Robustifying descriptor instability using fisher vectors," *TIP*, 2014.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.

[12] K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun, "Learning invariant features through topographic filter maps," in *CVPR*, 2009.

[13] M. Ranzato and G. E. Hinton, "Modeling pixel means and covariances using factorized third-order boltzmann machines," in *CVPR*, 2010.

[14] U. Schmidt and S. Roth, "Learning rotation-aware features: From invariant priors to equivariant descriptors," in *CVPR*, 2012.

[15] H. Cai, K. Mikolajczyk, and J. Matas, "Learning linear discriminant projections for dimensionality reduction of image descriptors," *TPAMI*, 2011.

[16] D. Gavrila and J. Giebel, "Virtual sample generation for template-based shape matching," in *CVPR*, 2001.

[17] Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *CVPR*, 2004.

[18] J. M. Morel and G. Yu, "Asift: A new framework for fully affine invariant image comparison," *SIAM J. Imaging Sciences*, 2009.

[19] P. Simard, Y. LeCun, J. S. Denker, and Victorri B., "Transformation invariance in pattern recognition-tangent distance and tangent propagation," in *Neural Networks: Tricks of the Trade*, 1996.

[20] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a mahalanobis metric from equivalence constraints," *JMLR*, 2005.

[21] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012.

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

[23] J. Geusebroek, G. Burghouts, and A. Smeulders, "The amsterdam library of object images," *IJCV*, 2005.

[24] I. Everts, J. C. van Gemert, and T. Gevers, "Per-patch descriptor selection using surface and scene properties," in *ECCV*, 2012.