DEEPPATCH: LEARNING INVARIANT LOCAL IMAGE DESCRIPTORS

Jiakai Zhang*, Jonathan Tompson* & Arjun Jain
* These authors contributed equally
Courant Institute of Mathematical Sciences
New York University
{zhjk, tompson, ajain}@cims.nyu.edu

ABSTRACT

This paper presents a convolutional network (ConvNet) based algorithm for efficient computation of robust local image descriptors. For an input image patch centered on a pixel, we learn an invariant low-dimensional non-linear embedding in feature space. Our descriptor does not rely on any scale or affine parameters from the detection stage and it is shown that our descriptor is highly invariant to a wide variety of geometric and photometric changes including scale, rotation, viewpoint change, image blur, complex illumination changes and compression artifacts. The proposed descriptor has been evaluated on the *pick-the-best-daisy* Winder et al. (2009) dataset and we show significant improvement over existing state-of-the-art descriptors, both in terms of speed and performance.

1 INTRODUCTION

Designing stable and accurate descriptors for local image structure is important for many computer vision tasks, such as (object) recognition, stitching image panoramas, wide baseline matching and tracking. In these tasks, a generic image descriptor is typically chosen which should be general enough to discriminate between all possible natural-image patches.

Image patch appearance is determined by combinations of material properties, such as color and texture, along with scene properties such as illumination conditions, viewpoint, scale, etc. A successful image descriptor should be able to satisfy two conflicting constraints: a) have sufficient discriminative power between material properties and semantic content for use in general vision tasks, while b) remaining invariant to transformations in the input space that obfuscates patch correspondence. Since invariance is inversely proportional to discriminative power Varma & Ray (2007), many state-of-the-art image descriptors have been designed Lowe (2004) or optimized Winder et al. (2009) to find a careful balance between these conflicting constraints.

The goal of our descriptor is to be invariant (i.e. constant) between local image patches corresponding to all projections of local 3D neighborhoods under changes in viewpoint, lighting, blur and compression artifacts. We also require that the descriptor be discriminative when the 3D neighborhood changes, even if the the difference in unnoticeable in the projected local patch. In an attempt to design a descriptor which encodes maximum invariance while still being highly discriminant, we propose a convolutional network (ConvNet) based architecture to learn an invariant, low-dimensional embedding in feature space. This paper makes the following contributions:

- To our best knowledge, for the first time we learn a model to create invariant descriptors starting from raw image pixels using supervised learning on a comprehensive dataset rather than by hand-crafting and engineering invariance into the descriptor analytically.
- The ConvNet architecture used to infer descriptor value from a local image patch is efficient; given a 64×64 and 32×32 pixel input patches, our implementation requires 1.8ms and 0.7ms respectively to compute its descriptor.
- Our descriptor is robust to photometric and geometric transformations and out performs state-of-the-art descriptors such as SIFT Lowe (2004), DAISY Tola et al. (2010) and Winder et al. Winder et al. (2009) on the *photo-tourism* Winder et al. (2009) dataset.

The paper is organized as follows. Section 2 describes related work. In Section 3, we explain our architecture for creating invariant feature space representations from image patches in detail. In Section 4 we empirically evaluate our descriptor, discuss it's properties and compare discriminative performance against other state-of-the-art descriptors such as as SIFT, DAISY and others.

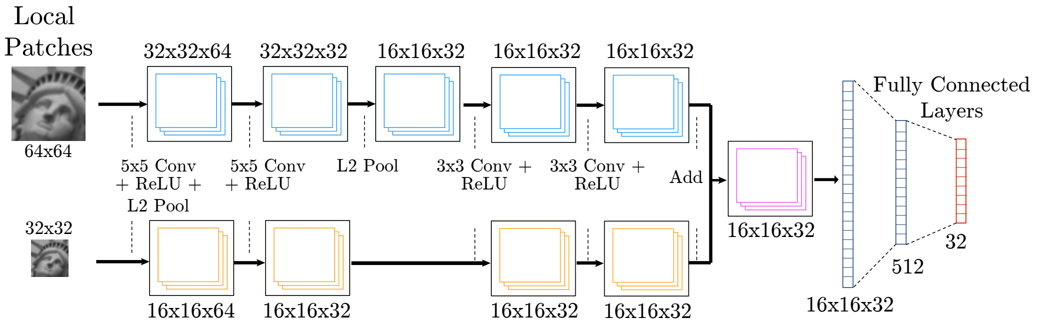


Figure 1: Multi-resolution Convolutional Network based Descriptor

2 Related Work

Designing an invariant local image patch descriptor is a fundamental problem in computer vision. Since a wide range of applications build on such descriptors, a long history of previous work can be found for this problem. For brevity we will only focus on recent prior-work that is most relevant.

By far one of the most popular and widely used patch-based descriptors is the Scale Invariant Feature Transform (SIFT) Lowe (2004). SIFT is a hand-crafted analytic function that uses the magnitude and orientation of local image gradients accumulated in sub-regions to build a feature histogram, which can be used to measure similarity between image patches. SIFT was later extended in the Gradient Location and Orientation Histogram (GLOH) method Mikolajczyk & Schmid (2005). Mikolajczyk and Schmid Mikolajczyk & Schmid (2005) demonstrate that SIFT and GLOH outperform other descriptors such as those obtained using shape context, steerable filters, spin images, differential invariants, complex filters, and moment invariants. Some other more recent descriptors include the Local Intensity Order Pattern (LIOP) Wang et al. (2011), KAZE Alcantarilla et al. (2012), Center-Symmetric Local Binary Pattern (CS-LBP) Heikkila et al. (2009), the shape of Maximally Stable Extremal Regions (MSER) Forssen & Lowe (2007), Robust Edge Aware Descriptor (READ) Rouzbeh Maani (2012), DAISY Tola et al. (2010) and its successor by Winder et al. Winder et al. (2009) where they learn the best parameters for DAISY in a supervised setting and show improved performance over DAISY.

Similar to SIFT, DAISY uses the magnitude and orientation of local gradients; however, the weighted sum of gradient orientation is replaced by the convolution of the gradient in specific directions with several Gaussian filters. Recently, it has been shown that the intensity ordinal information is more useful than the fixed location bins used by many descriptors such as SIFT and DAISY. The idea has been used by several descriptors such as LIOP Wang et al. (2011), MROGH Fan et al. (2011), and Multisupport Region Rotation and Intensity Monotonic Invariant Descriptor (MRRID) Fan et al. (2012). READ Rouzbeh Maani (2012) measures the similarity of the underlying structure to an edge and reports state-of-the-art performance till date.

Motivated by real-time applications, another line of effort is towards developing binary descriptors such as Binary Robust Independent Elementary Features (BRIEF) Calonder et al. (2010), Binary Robust Invariant Scalable Keypoints (BRISK) Leutenegger et al. (2011), Oriented Fast and Rotated BRIEF (ORB) Rublee et al. (2011) and Fast Retina Keypoint (FREAK) Alahi et al. (2012). A comparative evaluation of these descriptors is presented in Heinly et al. (2012). The recent paper by Miksik and Mikolajczyk Miksik & Mikolajczyk (2014) also compares some of these methods in the accuracy and speed trade-offs suggesting that binary descriptors provide comparable precision/recall results with SIFT and but perform better in terms of speed. On the other hand Miksik & Mikolajczyk (2014) reports that LIOP, MRRID, MROGH are slower but outperform SIFT and other binary descriptors.

3 Design of our Local Descriptor

The aim of this work is to learn an invariant yet discriminate mapping of input patch to a low dimensional descriptor. We want the descriptor space to be a metric space and therefore one should be able to measure the similarity between two descriptors by simply measuring euclidean distance. ConvNets have been shown to be endowed with high discriminative power and have been successful on a wide variety of computer vision problems Yaniv Taigman & Wolf (2014); Tompson et al. (2014); Szegedy et al. (2014), however to the best of our knowledge, they have not been used to learn a general-purpose invariant patch descriptor targeted at discriminating local patch correspondences.

In Section 3.1 we first present our ConvNet architecture for mapping high-dimensional input patches to a low-dimensional invariant representation. In Section 3.2 we show how we explicitly train this architecture to be invariant to geometric or photometric transformations.

3.1 Convolutional Network

Our multi-resolution ConvNet architecture is shown in Figure 1. It takes as input a 2 layer Gaussian pyramid created from a local 64×64 image patch, which is fed through two multi-resolution ConvNet banks. These convolution feature maps are then input to a 2 layer neural network, which then outputs a 32 dimensional feature descriptor.

The primary motivation for the use of multiple-resolution banks is efficiency. Empirically, we have found that we need convolution features with large spatial context in order to capture low frequency features of the input image patch. This can be performed either by using large convolutions in a single resolution bank or alternatively by using a constant (and small) size convolution kernel at multiple image scales. We employ the latter solution as it effectively reduces the number of trainable parameters (reducing over-training) and decreases computational load (since the number of operations for a spatial convolution is quadratic in the convolution size).

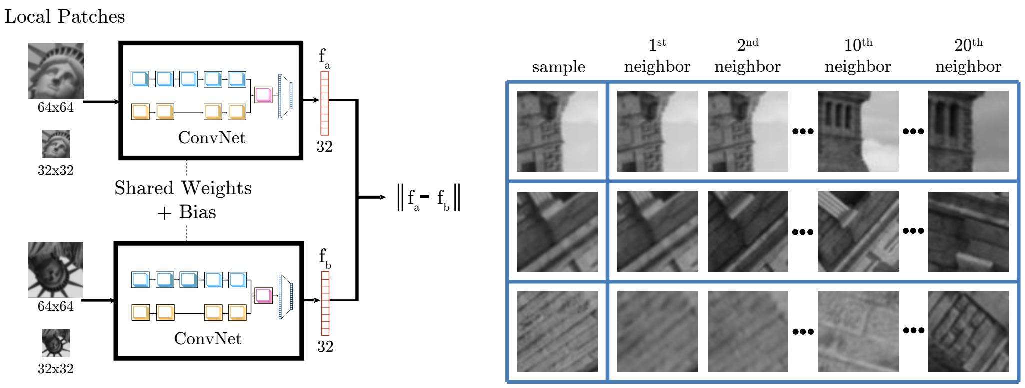
Each resolution bank is comprised of 4 stages of compound convolution-Non-Linearity-Pooling layers and we use a linear rectification unitNair & Hinton (2010) as our activation function for all intermediate non-linearity layers. The use of pooling here is important, firstly it adds a small amount of invariance to local input transformations (particular translations). Additionally, because it performs dimensionality reduction in the spatial domain, this reduces learning capacity in the first fully-connected network layer, which in turn prevents over-training.

Empirically we found that L_2 pooling (where local pooling regions are aggregated using a L_2 norm before decimation) output-performs Max-Pooling on this problem. We believe that this is because our feature descriptor performance benefits greatly when the ConvNet mapping from patch to output descriptor is as smooth as possible. Furthermore, max-pooling is tolerant to translations in input space but since we want to be able to discriminate between translation transformations, max-pooling is not a good choice for our problem.

3.2 Training Architecture

Training of the model in Figure 1 is carried out using the Siamese architecture shown in Figure 2a. It processes two image patches I_a and I_b through our ConvNet feature extractor from Section 3.1 to produce two 32 dimensional feature vectors f_a and f_b. During training, the weight and bias parameters of the two network branches are shared and updated in unison, so that the mapping from image to feature is consistent for both branches.

We then calculate the L_2 norm of the distance between feature vectors, which is then used as input to our objective function. Note that the use of L_2 norm here is not arbitrary. We want to design a feature descriptor where euclidean distances in output representation is meaningful, i.e. we want to design the training metric to be consistent to the targeted test-time metric.



(a) Network Architecture For Training

(b) Nearest-Neighbors in feature space

The Siamese network is trained using the DrLIM objective function proposed by Hadsell et al. Hadsell et al. (2006):

$$L(c) = \sum_{\{x_{i1}, x_{i2}\} \in P} L_i \left(c, \| f\left(c, x_{i1}\right) - f\left(c, x_{i2}\right) \| \right) \tag{1}$$

where c is the parameter vector of our ConvNet (weights and biases), $\{x_{i1}, x_{i2}\}$ is the ith image pair in the training set pairs P, $f(c, x)$ is the functional mapping $f \colon \mathbb{R}^{4096} \rightarrow \mathbb{R}^{32}$ defined by our ConvNet and L_i is the loss for the ith sample pair defined as:

$$L_i\left(c, x\right) = \begin{cases} \frac{1}{2} x^2 & \text{positive pair} \\ \frac{1}{2} \max^2\left(0, m - x\right) & \text{negative pair} \end{cases} \tag{2}$$

where m is a scalar margin. During BPROP we seek the solution to the following minimization problem: $c^\star = \underset{c}{\operatorname{argmin}} \left(L\left(c\right) \right)$. Effectively, this loss function applies an attractive force to image pairs with a positive label (i.e. pairs that are in correspondence) and repels negative pairs apart that are within the margin distance m.

The advantage of using a DrLIM-like contrastive loss function on this dataset is two-fold. Firstly, since the positive attraction term attempts to map input patches that have undergone complex transformations in input space (rotations, scaling, view and lighting changes, etc) to the same point in our $\mathbb{R}^{32}$ space, it forces the ConvNet to learn a stable and invariant mapping.

Secondly, unlike training a binary classifier on this problem, the output representation has many degrees of freedom with which to describe the semantic similarity between any given input pair. At no point does the network make a hard (and potentially unstable) Boolean decision about correspondence which may lead to over-training. If desired, a binary classifier can be trained on top of our descriptor and we show empirical results for this in Section 4. However in Section 4 we will also show that since our objective function maintains that euclidean distances within our $\mathbb{R}^{32}$ descriptors is a meaningful measure, a L_2 norm comparison has sufficient discriminative power to correctly classify patch correspondences and beat previous state-of-the-art results.

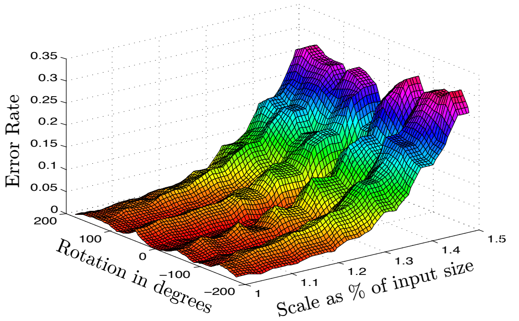
We use multiple data-augmentation strategies during training. This is done for two reasons: a) to prevent over-fitting and more importantly b) to encode invariance of sample input patches in descriptor space. At each epoch, we randomly scale ($s \in [1, 1.1]$), rotate ($r \in [-30°, +30°]$) and flip (with 0.5 probability) both image pairs together.

4 Experiments and Results

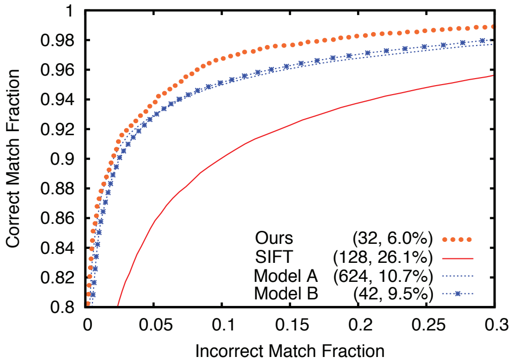
We compare our descriptor using the standard precision/recall measure on the photo-tourism Winder et al. (2009) datasets in Section 4.1.

In Figure 2b, we show the semantic similarity between ranked nearest-neighbors (measured using euclidean distance) in our $\mathbb{R}^{32}$ descriptor space for 3 sample patches from the liberty test-set. The 10th and 20th neighbors for the top row are very far apart in input space, but are multiple scales of the same region on the liberty structure, and is an example of our descriptors invariance to input scale. Similarly, the middle row shows that our descriptor is able to learn invariance to in-plane camera rotations, while the bottom row shows invariance to camera blur. Interestingly, the 10th neighbor for the bottom row is not from the same region of the building, but is semantically similar due to the presence of a brick-like texture showing that our descriptor is able to successfully encode textural information.

If f_i is the feature space representation for an input image patch p_i, in Figure 3a, we compute change in descriptor space resulting from the change in 2D rotation (R) and uniform scaling (S) of the input patch ($\frac{\partial f_i}{\partial S \partial R}$). We see that our descriptor is highly invariant to rotation to up to $\pm 30\%$. We also see that the invariance to scale is smoother and the descriptor only changes by $XX\%$ with even a $1.5\times$ change in scale. Note that $1.5\times$ is a large amount of scaling when dealing with local patches. We attribute this characteristic to our aggressive data-augmentation strategies performed during training.



(a) Error rate versus scaling and rotation of input samples.



(b) ROC curves for best models from Winder et al. (2009) and our model. Figures in brackets show error rates.

4.1 Evaluation on Photo-tourism Dataset

We evaluate our model on the 3 datasets proposed by Winder et al. Winder et al. (2009); Yosemite, Liberty and Notre Dame. Each dataset contains 100,000 random patch pairs with 50% of these pairs being positive correspondences. Like Winder et al. (2009), we report test performance of our descriptor on Liberty and Notre Dame. When testing on Liberty, we train on Yosemite and Notre Dame, and when testing on Notre Dame we train on Liberty and Yosemite.

For each trained descriptor we computed ROC curves and obtained percentage ($\%$) error rates when 95% of all correct matches were obtained. As it can be seen from Table 1, our descriptor significantly outperforms Winder et al. (2009) by a large margin. We would also like the reader to note that Winder et al. (2009) is sensitive to the descriptor parameters, such as the number of PCA components, descriptor dimension, etc. On the other hand, our approach simply works *out-of-the-box*. Furthermore, Winder et al.'s Winder et al. (2009) best model for Liberty is different from their best model for Notre Dame, but we exhibit improved performance on both datasets using exactly the same model structure. Thus, our model is robust to test datasets and exhibits better generalization performance.

In Figure 3b, we compare our descriptor to two best models from Winder et al. (2009): Model A which has a 624 output feature dimension and achieves an error rate of 10.7% and Model B which is obtained after applying PCA to the feature vectors and has a 42 feature dimension and performs slightly better with an error rate of 9.5% on the Notre Dame dataset. Our model has a constant descriptor dimension of 32 and achieves an error rate of 6.03% for the same dataset.

Dataset	Model	Error	Dim.
Notre Dame	Best model 1 from Winder et al. (2009)	9.71	37
Notre Dame	Ours	**6.03**	32
Liberty	Best model 2 from Winder et al. (2009)	17.14	42
Liberty	Ours	**8.66**	32

Table 1: Error rates comparison of our model and Winder et al. (2009) for Liberty and Notre Dame datasets.

5 Conclusion

We presented a method for training a general purpose local image descriptor that is invariant to a wide variety of geometric and photometric transformations. When this descriptor is compared against the state-of-the-art descriptors on standardized datasets, it clearly comes out as the winner. Note, for now, we learn our model directly on the pixel intensity values and do not use any color information yet. This is due to the fact that our training dataset is devoid of color information. In future, we also plan to incorporate color information and expect to achieve higher discriminative power in our descriptor space. In the future we also plan to work on a *one-shot* model which would on redundant convolutions (e.g. if there is overlap in the input image patches or for applications requiring dense descriptors).

References

Alahi, A., Ortiz, R., and Vandergheynst, P. Freak: Fast retina keypoint. In *CVPR*, 2012.

Alcantarilla, P.F., Bartoli, A., and Davison, A.J. Kaze features. In *ECCV*, 2012.

Calonder, M., Lepetit, V., and Strecha, C. Brief: Binary robust independent elementary features. In *ECCV*, 2010.

Fan, B., Wu, F., and Hu, Z. Aggregating gradient distributions into intensity orders: A novel local image descriptor. In *CVPR*, 2011.

Fan, B., Wu, F., and Hu, Z. Rotationally invariant descriptors using intensity order pooling. In *PAMI*, 2012.

Forssen, P. and Lowe, D. Shape descriptors for maximally stable extremal regions. In *ICCV*, 2007.

Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

Heikkila, M., Pietikainen, M., and Schmid, C. Description of interest regions with local binary patterns. In *PR*, 2009.

Heinly, J., Dunn, E., and Frahm, J.M. Comparative evaluation of binary features. In *ECCV*, 2012.

Leutenegger, S., Chli, M., and Siegwart, R.Y. Brisk: Binary robust invariant scalable keypoints. In *ICCV*, 2011.

Lowe, David. Distinctive image features from scale-invariant keypoints. In *IJCV*, 2004.

Mikolajczyk, K. and Schmid, C. A performance evaluation of local descriptors. In *PAMI*, 2005.

Miksik, O. and Mikolajczyk, K. Robust edge aware descriptor for image matching. In *ACCV*, 2014.

Nair, Vinod and Hinton, Geoffrey E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010.

Rouzbeh Maani, Sanjay Kalra, Yee-Hong Yang. Evaluation of local detectors and descriptors for fast feature matching. In *ICPR*, 2012.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. Orb: an efficient alternative to sift or surf. In *ICCV*, 2011.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going Deeper with Convolutions. In *ArXiv e-prints*, 2014.

Tola, E., Lepetit, V., and Fua, P. Daisy: An efficient dense descriptor applied to wide-baseline stereo. In *PAMI*, 2010.

Tompson, J., Stein, M., LeCun, Y., and Perlin, K. Real-time continuous pose recovery of human hands using convolutional networks. In *TOG*, 2014.

Varma, M. and Ray, D. Learning the discriminative power-invariance trade-off. In *ICCV*, 2007.

Wang, Z., Fan, B., and Wu, F. Local intensity order pattern for feature description. In *ICCV*, 2011.

Winder, S., Hua, G., and Brown, M. Picking the best daisy. In *CVPR*, 2009.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato and Wolf, Lior. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.