CDIKP: A Highly-Compact Local Feature Descriptor

Yun-Ta Tsai, Quan Wang, Suya You
Computer Science Department
University of Southern California
{yuntatsa, quanwang,suya}@usc.edu

Abstract

A new feature descriptor is presented for object and scene recognition. The new approach, called CDIKP, uniquely combines the scale-invariant feature detection with a robust projection kernel technique to produce highly efficient feature representation. The produced feature descriptors are highly-compact in comparisons to the state-of-the-art, do not require any pre-training step, and show superior advantages in terms of distinctiveness, robustness to occlusions, invariance to scale, and tolerance of geometric distortions. We extensively evaluated the effectiveness of the new approach with various datasets acquired under varying circumstances.

1. Introduction

Local features with invariant descriptions are important for many tasks in computer vision and object recognition, including object and scene recognition, wide baseline image matching and registration, and content-based image retrieval. In a feature based approach, the images are transferred into a set of highly-distinctive local features, each representing a quasi-independent salient region of object within the scene. The features are represented with robust visual descriptors to capture both structural and appearance properties of scenes that are largely invariant to imaging and viewing variations.

There are a number of local features and invariant descriptors that have been proposed for various visual recognition tasks. Features can be extracted in single or multiple scale(s), and either densely, at random, or sparsely according to local informativeness of scenes. For example, the Harris detector that is based on the auto-correlation of image is often used for finding local features. Mikolajczyk and Schmid [1] combined the Harris detector with automatic scale selection to detect distinctive feature that is scale-invariant. Lowe [2] proposed to extract salient features using local 3D peaks detected in a Difference-of-Gaussian (DoG) pyramid. There are a number of other state-of-the-art detectors that find scale- or affine-invariant local features and salient regions [3].

Many methods for feature descriptions have been suggested, which can incorporate various degrees of resistance to common perturbations such as viewpoint changes, geometric deformations, and photometric transformations. Among the approaches, the SIFT descriptor has been shown to outperform other descriptors [4]. The SIFT descriptor is based on the gradient distribution in salient region, and constructed from a 3D histogram of gradient locations and orientations. A 128-dimension vector representing the bins of the oriented gradient histogram is used as descriptor of salient feature [2].

However, the high dimensionality of SIFT descriptor is a significant drawback, especially for online or large-scale dataset applications. For a typical outdoor scene, for example, the SIFT usually produces several hundreds of local features, yielding a large high-dimensional feature space needs to be searched, indexed, and matched.

Several researchers have addressed the problem of dimensionality reduction for feature descriptors. For example, Herbert et al [5] proposed an approach (SURF) that combined the Hessian matrix-based measure for the detector and Haar-wavelet responses for the descriptor, resulting in a 64-dimension feature representation. PCA-SIFT proposed in [6] reduced the dimensionality of descriptor to the range of 36, while remaining a comparative performance to the original SIFT. The key of PCA-SIFT is to apply the standard Principal Components Analysis technique to the gradient patches extracted around local features, therefore yielding a compact feature representation. However, the PCA-SIFT needs an offline stage to train and estimate the covariance matrix used for PCA projection. This typically requires the system to collect and train a large, diverse collection of images *prior* to use, (it often needs to re-train and re-estimate the covariance matrix when the image database is expanded or the scenes have significant changes), thereby impeding its widespread use and benefits.

This paper presents our efforts in developing an efficient local feature and its invariant descriptor for scene recognition. Our main contributions lie in a novel approach that uniquely combines the scale-invariant feature detection with a robust kernel-based representation technique to produce highly efficient feature representation. We named the approach Compact Descriptor through Invariant Kernel Projection (CDIKP). The produced feature descriptors are highly-compact (20-Dimension) in comparisons to the state-of-the-art (e.g. SIFT: 128-D, SURF: 64-D, and PCA_SIFT: 36-D), do not require any pre-training step, and show superior advantages in terms of distinctiveness, robustness to occlusions, invariance to scale, and tolerance of geometric distortions.

2. Approach

Figure 1 depicts the main steps of our approach, which are detailed in following sections.

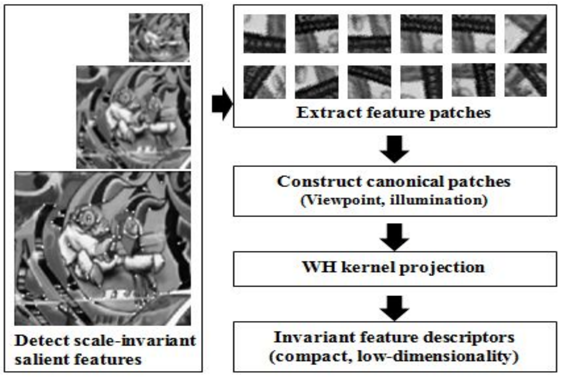

Extract feature patches
Construct canonical patches (Viewpoint, illumination)
WH kernel projection
Invariant feature descriptors (compact, low-dimensionality)
Detect scale-invariant salient features

Fig 1: Algorithm structure

2.1. Scale-invariant Feature Detector

Our approach selects multi-scale salient features/regions with the scale-invariant detector [2] where 3D peaks are detected in a DoG scale-space. The peaks in a DoG pyramid have been shown to provide the most stable interest regions when compared to a range of other interest point detectors.

Three spatial filters are used in the detector. First, a high frequency-passed filter is employed to detect all the candidate features with local maximum responds in the DoG pyramid. The second filter is a distinctiveness filter that removes the unstable features usually lying along the object edges or linear contours. The third filter is an interpolation filter that iteratively refines the feature locations to sub-pixel accuracy. Finally, the dominant orientation and scale are computed and assigned to each detected feature. The dominant orientation and scale will be used for view normalization to achieve viewpoint invariant. Figure 2 shows an example of the detected salient features.

2.2 Kernel-based Feature Descriptor

Discrimination power is an important factor required for object recognition with high data variability. We base our feature descriptor on a projection kernel scheme, because the projection kernel techniques have demonstrated strong discrimination performance and they are well-established analytical tools that are useful in variety of contexts including discriminative classification, scene recognition and categorization. Another attractive feature of the projection kernel techniques is their innate data compaction that can efficiently map high-dimensional data to a compact representation with much lower dimensionality. This is a very attractive property for image description from which we could produce compact, lower-dimensional descriptors.

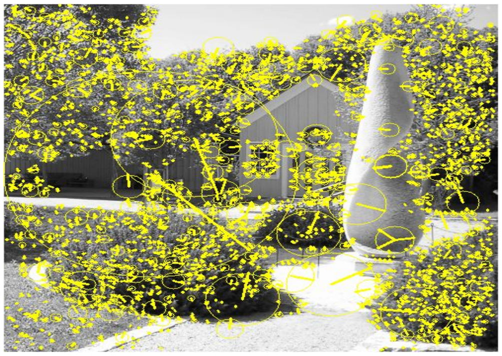


Fig 2: Detected salient features with feature locations, dominant orientations, and scales

Choosing an appropriate kernel function is a key for efficient projection kernel schemes. Two important factors have to be considered: the kernel functions should be ordered in a way such that the lower bound becomes tight after only a small number of projections, and the kernels should be efficient to enable fast computation [7, 8].

2.2.1 Walsh-Hadamard Kernel

We decided to use Walsh-Hadamard (WH) kernel because of its good performances in discrimination and computational efficiency. Mathematically, the WH kernel vectors can be recursively constructed as

$$H_k = \begin{bmatrix} H_{k-1} & H_{k-1} \\ H_{k-1} & -H_{k-1} \end{bmatrix}, \; H_0 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Therefore, the elements of WH kernel vectors are orthogonal and made up entirely of 1 and -1. Computation of the WH transform involves only integer additions and subtractions. Given an image patch of size $k \times k$, its WH transform is computed by projecting the patch onto k^2 WH kernel vectors. It has been shown in [7] and also confirmed by our experiments that the first few WH projection vectors can capture a high proportion of information contained in the image (Figure 3). These unique properties of WH kernel projection lead us an efficient tool to build compact descriptors.

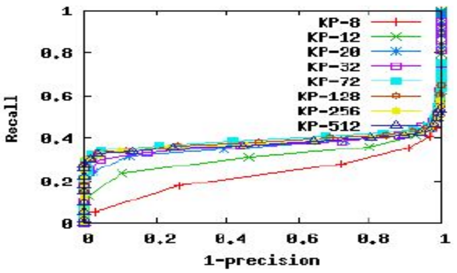


Fig 3: Impact of the lengths of WH projection vectors on feature matching. Note: KP-8 means a 8-dimensional feature descriptor, in which the first 4 projection vectors of WH transform are extracted separately for x- and y-component of an image patch, and then combine them to produce the 8-dimensional descriptor

2.2.2 Generate Descriptors with WH Projections

WH kernel projection, however, does not naturally have the important property of geometry invariance, thus it cannot handle geometric distortions caused by viewpoint or pose changes. We solve this problem by performing a viewpoint normalization step on the basis of the feature's dominant orientation and scale.

Constructing the canonical views of features is relatively simple and fast. We first extract local patches centered at the feature locations from the Gaussian pyramid constructed in the above step of feature detection. The size of patch varies with the scale at which the feature was detected. Under the assumption of local planarity, a new canonical view of the local patch (with fixed size and scale) is synthesized by image warping with the feature's dominant orientation and scale. This corresponds to a regular re-sampling process in an affine space. Note that the size of the canonical patch is fixed and has to be in the power of 2, as required by WH transform. Our extensive experiments show that the size of 32×32 gives the optimal results (Figure 4).

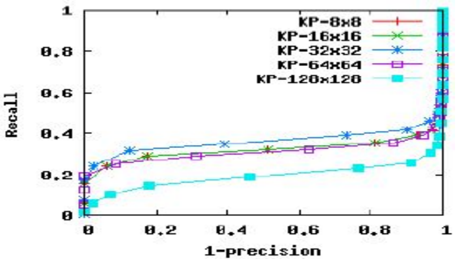
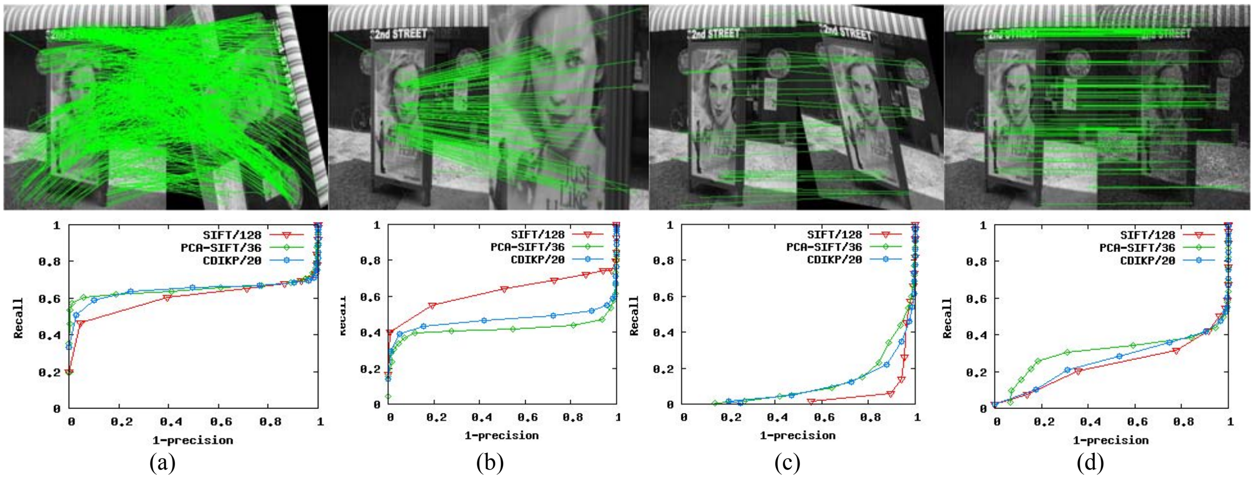


Fig 4: Impact of patch sizes on feature match performance

To reduce the effect of photometric changes, we use gradient for each patch in WH transform. We have evaluated several gradient computation and forms, and found that the Gaussian weighted first-order derivatives of pixel intensity along horizontal and vertical directions seemed to yield the most robust results to compensate the substantial changes of illumination. Thus, we first calculate the first-order derivatives in x and y directions within a local patch, and then weight the directional derivatives using a weighted Gaussian kernel: $Gauss\,(\sigma, d) \times W(d)$, where $W(d) = (1-d)/(\frac{S}{2})$, $\sigma = S/2$, d is the relative distance between current pixel to the center of the patch, and S is the patch size. In this way, we obtain a pair of gradient maps v_x and v_y for each local patch which is canonically normalized to viewpoint and photometric variances.

We then use the WH kernel projection to extract significant components contained in the local patches to generate feature descriptors. Since we obtain two 1024-element gradient maps v_x and v_y for each patch/feature, we apply the WH transform twice to the gradient maps: one for v_x and one for v_y. Finally, the first 10 projection vectors of each WH transform are extracted and combined to produce a 20-dimension feature descriptor that is compact, distinctive, and viewpoint and illumination invariant.



(a) (b) (c) (d)

Fig 5: Performance evaluation under different imaging and viewpoint variances

3. Experimental Results

We have extensively tested and evaluated the proposed approach using various datasets including synthesized data, a standard evaluation set, and our own datasets acquired under varying circumstances. We evaluated the effectiveness of our approach, in comparison to other descriptors, in the terms of distinctiveness, robustness and invariance. Due to space limitation, we cannot present all the data in this paper. The extensive results will be provided upon request. We are also planning to release the algorithm code for public evaluation.

3.1 Synthesized Data Evaluation

We collected a dataset of images and intentionally distorted them with various geometric and photometric transformations. For a pair of test images, we ran the CDIKP algorithm to automatically select distinctive features, generate descriptors, and find the feature matches. The results were evaluated using the metric of recall-precision graphs, as in [4, 5, 6]. We conducted performance comparisons to standard SIFT, and PCA-SIFT. In our tests, we tried to use the same set of parameters (e.g. pyramid levels, filter sizes, and thresholds) for all the three methods.

Figure 5 shows results of the CDIKP approach to a scene under different distortions, where (a) is the matched features for the original image being rotated 70 degree; (b) for 250% scaling; (c) for 0.4-x, 0.1-y shearing; and (d) for 100% illumination change and adding 15% Gaussian noise.

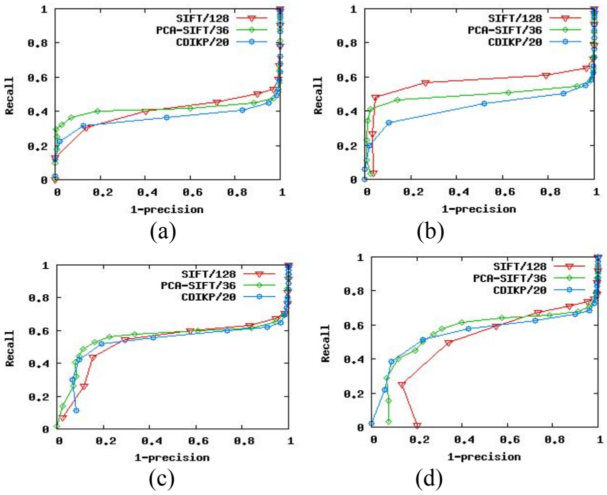
3.2. Standard Test Dataset with Ground Truth

We tested our approach using the INRIA dataset [4]. These are images of real scenes with recovered deformation parameters used as test ground. Figure 6 shows the results for several cases: (a) rotation and scale (Boat), (b) viewpoint changes (Wall), (c) image blur (Bikes), and (d) lighting changes (Leuven).

We can see from these results that the proposed CDIKP descriptor remains a very comparative performance, sometimes outperforms SIFT in recall for the same level of precision. Meanwhile, it is more compact and efficient to compute.

3.3. Scene Recognition Application

We used the approach for object recognition application intending to the content-based image retrieval (CBIR) on mobile-platform. Figure 7 demonstrates a scenario of applying the approach to automatically localize and recognize various commercial logos in nature environments. Note that the nature images were captured with a digital camera, and the logo images were downloaded online from the company websites. The logo images are usually computer-generated graphics that have significant differences in geometry and colors with real images. These examples demonstrate the value of the proposed approach for mobile multimedia applications such as product advertising and shopping.



(a) (b)

(c) (d)

Fig 6: Evaluation with INRIA test dataset [4]

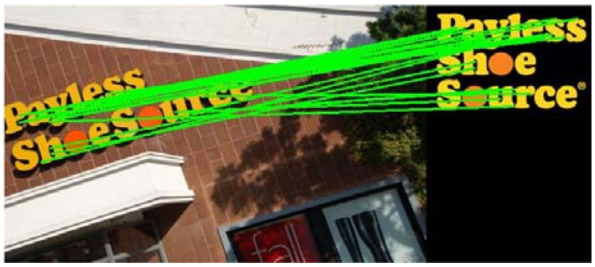


Fig 7: Apply CDIKP to localize and recognize commercial logos in nature environments

4. Reference

[1] K. Mikolajczyk and C. Schmid, “Scale and affine invariant interest point detectors,” Int’l J. of Computer Vision, 60(1): 63–86, 2004

[2] D. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” Int’l J. Computer Vision, 2004

[3] C. Schmid, R. Mohr, and C. Bauckhage, “Evaluation of interest point detectors,” Int’l Computer Vision, 37(2):151–172, 2000

[4] K. Mikolajczyk and C. Schmid, “A Performance Evaluation of Local Descriptors,” IEEE Trans. on PAMI, 27(10): 1615-1630, 2005

[5] H. Bay, T. Tuytelaars and L. V. Gool, “SURF: Speeded Up Robust Features,” the 9th European Conference on Computer Vision, 3951”404-417, 2006

[6] Y. Ke and R. Sukthankar, “PCA-SIFT: A more distinctive representation for local image descriptors. CVPR, 2:506-513, 2004

[7] Y. Hel-Or and H. Hel-Or, “Real-Time Pattern Matching Using Projection Kernels,” IEEE Trans. on PAMI, 27(9): 1430-1445, 2005

[8] Q. Wang and S. You, “Real-Time Image Matching Based on Multiple View Kernel Projection,” in proceedings of CVPR, 1-8, 2007